2020-03

Machine learning models for predicting decisions to be made by small scale dairy farmers in Eastern Africa

Mwanga, Gladness George

NM-AIST

MACHINE LEARNING MODELS FOR PREDICTING DECISIONS TO BE MADE BY SMALL SCALE DAIRY FARMERS IN EASTERN AFRICA

Gladness George Mwanga

A Dissertation Submitted in Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy in Information and Communication Science and Engineering of the Nelson Mandela African Institution of Science and Technology

Arusha, Tanzania

March, 2020

ABSTRACT

In dairy, lack of decision support tools for identifying farmers' needs and demands have caused many programs, strategies, and projects to fail. This has led to the inefficient and fragmented allocation of scarce development resources. This study demonstrated how machine learning (ML) can be used as a tool to bridge this gap; by developing ML models to be used in identifying factors that can influence farmers decisions, predicting decision to be made by a farmer and forecast on farmers demands regarding to their specific need or service. Four countries: Ethiopia, Kenya, Tanzania and Uganda were selected for this study.

In the course of the study four models were developed one for each country with regard to the usage of animal supplements, keeping of exotic animals, use of Artificial insemination (AI) as breeding methods and animal milk productivity. Data was collected through face to face interviews, from a total of 16 308 small scale dairy farmers in Ethiopia (n = 4679), Kenya (n = 5278), Tanzania (3500) and Uganda (n = 2851). The decision tree algorithm was used to model categorical problems (use of supplement and breeding decision), which attained the accuracy of 78%-90%. Moreover, K-nearest neighbor was employed for numeric problems (keeping of exotic animals and animal milk productivity) with an accuracy of 0.78-0.96 Adjusted R^2 value.

The use of ML techniques assisted in classifying farmers based on their characteristics and it was possible to identify the key factors that can be taken then prioritized to improve the dairy sector among countries. Also, the results of this study offer a number of practical implications for the dairy industry where the proposed ML models can enable decision-makers in developing the National Dairy Master Plan and design policies that promote the growth of smallholder dairy farming. Moreover, these models shade light to potential service providers and investors who want to invest in dairy to identify potential areas or groups of farmers to focus with.

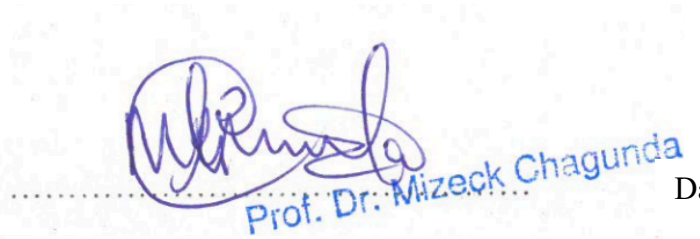
AUTHOR'S DECLARATION

I, Gladness George Mwanga do hereby declare to the Senate of Nelson Mandela African Institution of Science and Technology that this dissertation is my own original work and that it has neither been submitted nor being concurrently submitted for degree award in any other institution.

…….. ……

Gladness George Mwanga Date

The above declaration is confirmed



…… Date 23/03/2020

Professor Dr. Sc. Agr. Mizeck Chagunda

…… Date 23/03/2020

Eng. Dr. Zaipuna O. Yonah

…… Date 23/03/2020

Dr. Mussa Ally

COPYRIGHT

CERTIFICATION

The undersigned certify that they have read and hereby recommend for submission to the Nelson Mandela Institution of Science and Technology (NM-AIST) a dissertation titled Machine learning models for predicting decisions to be made by small scale dairy farmers in Eastern Africa, in fulfillment of the requirements for the degree of Doctor of Philosophy in Information and Communication Science and Engineering of the Nelson Mandela African Institution of Science and Technology.

Date 23/03/2020

Professor Dr. Sc. Agr. Mizeck Chagunda

Date 23/03/2020

Eng. Dr. Zaipuna O. Yonah

Date 23/03/2020

Dr. Mussa Ally

ACKNOWLEDGEMENTS

First and foremost, I give honor and glory to God for his unmerited favor and for his grace and love.

Second, I would like to express my profound gratitude to all my advisors Prof. Mizeck Chagunda, Dr. Denis Mujibi, Eng. Dr. Zaipuna O. Yonah, Dr. Svetlana Lockwood and Dr. Mussa Ally for their continuous support during my PhD study. I could not make this far without their support, motivation, leadership, and immense knowledge. I could not have imagined having a better team of advisors and mentors for my PhD study. I was very privileged to meet Prof. Mizeck and Dr. Denis, they inspired me, being a role model and helped me to grasp my subject better. I greatly value their advice and suggestions they provided throughout my research.

Besides my advisors, I would like to thank Prof. Morris Agaba for taking the initiative to mentor me in this interdisciplinary research. I value his support.

In addition to academic support, I extend my special thanks to PEHPL project and Nelson Mandela African Institution of Science and Technology (NM-AIST) for funding and granting me the opportunity to pursue my PhD studies. Furthermore, the Programme for Emerging Agricultural Research Leaders (PEARL) project for allowing me to grow as a researcher and for granting me access to use the project data for this research.

I also thank Prof. Guy Palmer for allowing me to visit Washington State University (WSU) and doing everything possible; paving a way and to enable my stay and my studies a success while being at WSU. Without forgetting all the support, I got from Prof. Mizeck who made my trip to Hohenheim University useful. I was fortunate to meet and interact with a number of colleagues at Hohenheim.

My deepest gratitude goes to all my friends and the people that I have been working together on various research projects.

Lastly, I would like to thank my family for everything they have done for me. Your love and encouragement means a lot. I thank my lovely husband, and my two daughters for being so supportive, encouraging, and patient throughout the course of my study program. And most of all my parents and my brothers for their endless love. Thank you.

DEDICATION

I dedicate this project to God Almighty. I also dedicate this work to my lovely husband Timothy Yusto Wikedzi and my two daughters Eliora Wikedzi and Elspeth Wikedzi. I could not, and most certainly would not have done it without your endless love and support.

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AI	Artificial insemination
AI	Artificial intelligence
ATM	Automated Teller Machine
AU-IBAR	African Union Inter-African Bureau for Animal Resources
AUC	Area under the Curve
AUROC	Area under the Receiver Operating Characteristic
BR	Boruta
CPU	Central processing unit
CV	Cross Validation
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DNA	Deoxyribonucleic acid
DT	Decision tree
EU	European Union
FAO	Food and Agriculture Organization
FHS	Framework for household system
GCC	Gulf Cooperation Council
GHZ	Gigahertz
GMM	Gaussian mixture model
GPS	Global Positioning System
ID	Identification
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things
KNN	K-nearest neighbor
LR	Logistic Regression
ML	Machine learning
NAGRIC	National Animal Genetic Resources Centre
NAIC	National Artificial Insemination Centre
ODK	Open Data Kit
QC	Quality control
RAM	Random-access memory
RCO	Receiver Operating Characteristics
RF	Random forest
RFID	Radio-frequency identification
ROC	Receiver Operating Characteristic
SAS	Analytics Software Solutions
SCC	Somatic cell count
SD	Standard deviation
SSA	Sub-Saharan Africa
SVM	Simple vector machine
TDCU	Tanga Dairy Cooperative Union
UGX	Uganda Shilling Exchange
USD	United States Dollar

VUI	Voice user interface
WTO	World Trade Organization

CHAPTER ONE

INTRODUCTION

1.1 Background of the study

Livestock agriculture plays many different roles in supporting families. It has been the source of food, income, asset saving, employment and wellbeing of most rural households. It also plays a significant role in alleviating poverty of over a billion people in the world. Milk and dairy products account for about 14% of global agricultural trade (FAO, 2016). In 2013, it was reported that dairy worth USD 328 billion in terms of liters (770 billion liters) produced globally and is expected to grow to 177 million tons of milk by 2025 (FAO, 2016). In this global worth, the developing world including Africa is ranked second with a greater number of dairy animals as it maintains two-thirds of the total herd in the world (Fig.1). However, currently, Africa and the developing world are not ranked even in the top ten lists of milk producers in the world (Fig. 2). Instead, they are the leading importers of milk products from developed countries (Fig. 3). This trend largely reflects an increase in livestock numbers, rather than productivity gains.

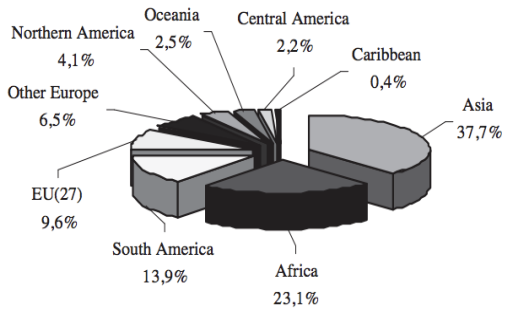


Figure 1: Distribution of the world's dairy cows in 2009 (Bulletin of the IDF, 2010)

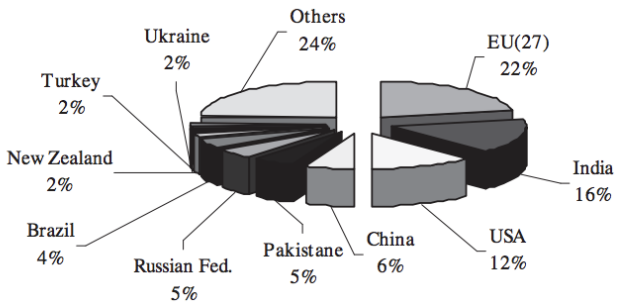


Figure 2: Distribution of world milk production (697 million tons) in 2009 (FAO, 2011)

In the developing world, it is estimated that 80 to 90 percent of milk is produced in small-scale farming systems (FAO, 2016). These operations are based on low inputs, and small heard size. For example, Mwanga, Mujibi, Yonah and Chagunda (2018) recently reported a herd size of between 1 and 13 cows per farm in Eastern Africa. Also, production per dairy animal is still low. Even for countries that are considered to be the base for dairy sites their animal production varies from 850-3150 liters/cow equivalent to a net farm income of $294 per year (Richards *et al.,* 2015). This figure is too low in comparison to milking cows in more intensive dairy farms

of developed countries where most of the producers are large scale farmers and on average one animal can produce up to 7800 liters/cow per year (VanLeeuwen *et al.,* 2012).

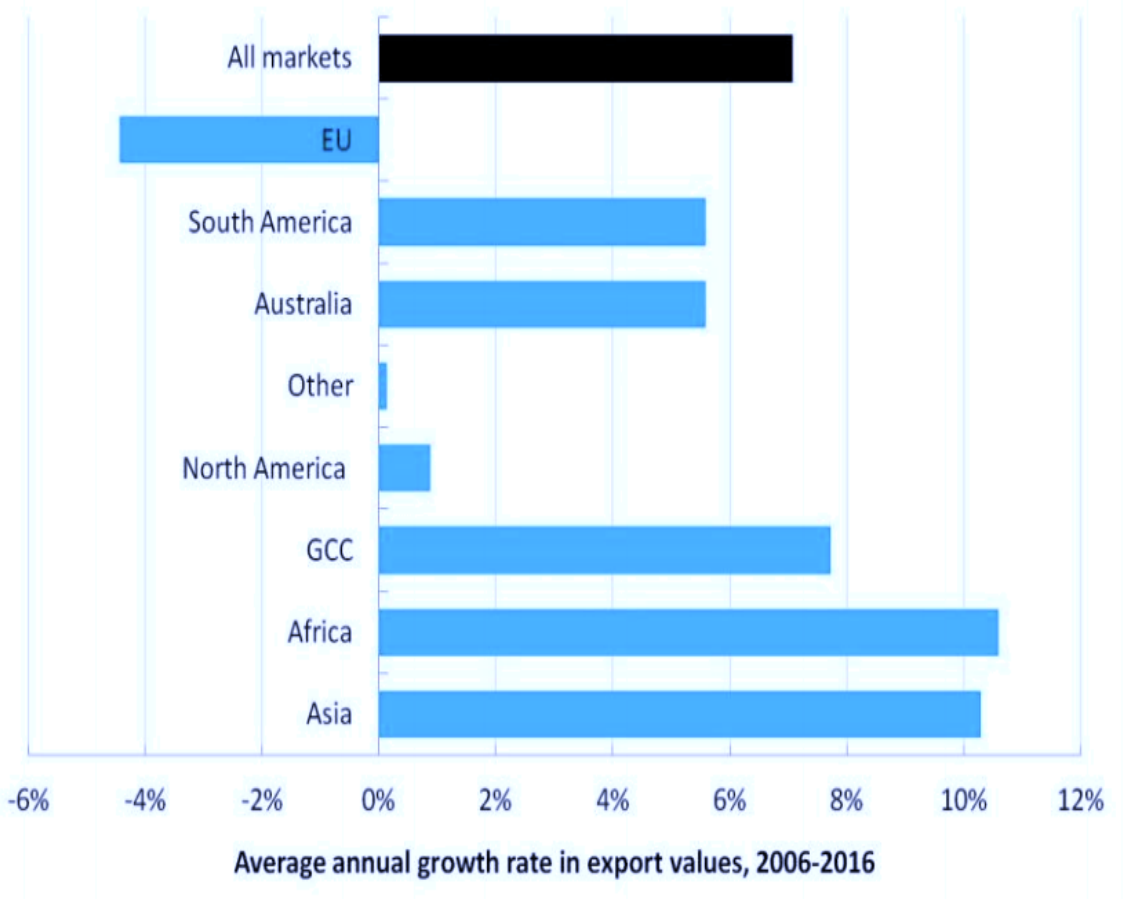


Figure 3: Average annual growth rate in dairy exports for New Zealand; From 2006-2016

Learning from other successful countries, this gap has been contributed by various factors, which are similar across different developing countries (Kanui & Ikusya, 2016). These include, among others: animal diseases, low adoption of improved technologies, and poor animal husbandry practices e.g. keeping of inappropriate cattle breeds. Moreover, the technological advancements of most dairy farms for developed countries have contributed a lot in improving their productivity. i.e. the use of ML, sensors technology and robotics have facilitated their decision-making process and automation of animal management practices. Where a farmer is able to detect diseases before it occurs, planning, and reducing the running costs by hiring a few workers.

In Sub-Saharan, small scale dairy farmers are considered as the main producer. They contribute more than 60% of the total milk produced. Due to the increase in milk demand which is also expected to double in Sub-Saharan Africa, farmers are required to increase their production

(The World Bank, 2014). Hence, a better support system to facilitate production is essential especially to small scale farmers who are the main producers (Richards *et al.,* 2015).

1.2 Problem statement

The success of many small-scale farms depends on public subsidized services. Though, the dairy sector has been losing millions of dollars every year trying to implement different technologies, services, and strategies to be adopted by farmers. Most of these initiatives have been receiving little attention from farmers due to a lack of analytical tools to scrutinize farmers' preferences, needs, demand and identifying factors that influence their decisions. For example, for the year 2017-2021, the government of Tanzania is planning to invest USD 101 million to improve the dairy sector in terms of animal feed, breeding technologies, animal health, and marketing. But the government continues to report the low adoption of these technologies and services by farmers. Also, in the year 2015, only 26% adopted breeding strategies 35% health services, and 50% adopted some feeding strategies. This is a common case for most SSA countries (Mugisha, Kayiizi, Owiny & Mburu, 2014; Tefera, Lagat & Bett, 2014) . The reasons for the low uptake of these technologies by farmers have never been clearly established across the main dairying countries in Africa. Therefore, understanding the key drivers of a farmer's choice is critical if the adoption rates are to be increased.

Moreover, various interventions and policies which are defined by a set of strategies and initiatives are being documented regularly (Ministry of Agriculture, 2016; Mbwambo, Nigussie & Stapleton*,* 2017; Morgan, 2018). These strategies can be set for a specific time frame and have to be prepared, planned and allocated budgets for implementation. A good strategy will need to consider farmers' preferences while considering existing barriers and resources. Hence, knowing farmers' preferences has become important. Also, in setting up strategies, a number of components need to be considered that require decision-makers to prioritize these strategies (Morgan, 2018). Usually, the list of priorities changes or may need to be revised from time to time in order to accommodate various changes happening on the farm. Similarly, due to limited resources quantification of resources needed is even critical.

It has been acknowledged that knowing farmers' demands and preferences assists decision-makers to properly allocate the right resources needed at a time (Hansson & Lagerkvist, 2016). Also, it lays the ground for other decision-makers such as inputs suppliers and other investors to identify potential business sites thus help to reduce risks and uncertainties. It was reported

that due to lack of information many programs and projects have been poorly designed and inadequately targeted which has often led to the inefficient and fragmented allocation of scarce development resources (Ugo Pica-Ciamarra *et al.,* 2014). Therefore, the availability of a decision support system to guide decision-makers throughout the process becomes vital. However, these systems have been lacking to this middle groups (peoples who support farmers) including service providers, policymakers, traders, extension workers and other stakeholders (Dudafa, 2013). This has led to the absence of clear roadmaps to develop the livestock sector, which persistently has hindered productivity.

1.3 Rationale of the study

Machine Learning (ML) techniques have been widely used in developing decision-support tools to get the insights needed to make better decisions. Based on its functionality that it continuously assesses and learns from data, to identify various patterns. Machine Learning is being increasingly used in different sectors including financial, health care, retails and social media (Khare, Jeon, Sethi & Xu, 2017). In dairy production, it has been mostly used in disease detection and surveillance (Yazdanbakhsh, Zhou & Dick, 2017), estrous detection (Shahriar *et al.,* 2016), animal behavior monitoring (Benaissa *et al*., 2017), and animal traceability (Rahman *et al.,* 2018). In business, ML is now commonly used as a business-decision making tool. Where it assists clients to derive meaningful business insights from their customers' data, i.e. predicting customer behaviors, purchasing patterns, customer segmentation and predicting their lifetime customers (Yeomans, 2015). Moreover, it has been well adopted in health with the capability of automating various decision-making processes such as identifying high-risk patients, recommending medicines for patients, predicting readmissions, to mention but a few.

In the livestock sector, the use of ML algorithms to support policy decision-making is still in its infancy stage. Similarly, there is still a lack of research probing how the use of these technologies' influences policymakers' and other livestock stakeholders' in decision-making practices. While ML is currently being used for other systems there is potential for the technology to do much more in the livestock sector.

The new trend is to introduce the use of analytical decision-supporting tools implemented using ML technologies as an approach to facilitate evidence-based decision making by livestock stakeholders. The advantage of using ML-based decision supporting systems is the automation of processes. Where information is automatically extracted from data by simply training the

model with data. Thus, the user does not need to spend a lot of time and labor in constructing and maintaining the system. Also, ML helps to capture real-world patterns much better. In the real world, decisions can be influenced by several factors that cannot be captured by linear models. A non-linear predictive model can flexibly capture the relationship between variables and outcomes, and hence, usually predict better. Moreover, the recommendation may be better when using the non-linear form of knowledge, compared to a standard recommendation that relays on correlation with the outcome variable.

Therefore, this study aimed at developing ML models that can be used in identifying key drivers that influence farmers' decisions. Also, to use the ML models to be able to predict farmers' demands in regard to farm inputs and services. The expectation is that the developed models would assist policymakers, service providers, and other stakeholders to identify farmers' preferences, identify the area of attrition, and discovery appropriate solution that can suit farmers.

1.4 Objectives

To answer our research questions, the following research objectives were pursued:

(i) To Characterize the farmers' decision-making process in order to identify patterns of information that influence farmers' decisions. Also, to perform cross-validation to identify if there are dynamics across the regions.
(ii) To perform features engineering and model selection for identifying appropriate predictors and ML algorithms to be used for models' development.
(iii) To test the algorithms used for model development in a different environment.

1.5 Research questions

To better understand how farmers, make decisions and be able to model, the research work was guided by the following research questions

(i) (a) What factors influence/drive farmers' decisions?
    (b) Are drivers influencing farmers' decisions cut across the regions?
(ii) (a) What are the best predictors, and ML algorithms to be used in developing robust predictive models?

(b) How can models be designed to accurately identify factors influencing farmers' decisions and accurately predict the decisions to be made by farmers?

(iii) Are the developed models robust for predictions and how will ML algorithms behave in a different environment?

1.6 Significance of the study

Presented in this chapter, is the need of having decision-supporting tools to guide livestock stakeholders including policymakers, services providers, and other stakeholders to make informed decisions. This research, therefore, aimed at offering ML decision support tools that can be used in identifying factors influencing farmers' decisions but also being able to predict decisions to be made by a farmer given a set of factors. To achieve the main goal the study deployed ML techniques. The study started by reviewing how machine learning has been used in dairy sectors and its value in improving productivity. Next, farmers were characterized to understand their characteristics and explore various factors that can influence their decisions for selecting a set of features to be used in model development. Then all hypothesized predictors were screened to identify key predictors that were used for the model development. This process involved two process features engineering and model selection. The last step was to evaluate the selected model using a new set of data.

In conclusion, the research will help to give insight and guide decision making process of different livestock stakeholders. Thus, the implications of our research will be to:

(i) Have a clear understanding of why farmers make certain decisions and what constrain farmers decisions in order to meet their needs and preferences

(ii) Identifying key drivers that contribute to the dairy productivity for prioritization of different initiatives and strategies

(iii) Predicting farmers' demands with regard to various services for proper allocation of resources

(iv) Assist policymakers in planning and setting up strategies

1.7 Delineation of the study

This research is concerned with the use of ML to facilitate decision-making process in the livestock sector. Machine learning can be considered as a branch of artificial intelligence based on the idea that the system/computer can learn from data to identify patterns and make decisions with minimal human intervention without being explicitly programmed. The algorithms use statistical analysis to predict output and it keeps on updating outputs as new data becomes available. Machine learning instructions are not directly provided by the programmer; thus, the aim of this study was to construct ML models to fits the given data. The programs designed had to perform a repetitive process of feature and model's selection and by modifying various algorithms parameters to obtain a robust model. Models development in this study are defined as taking data collected from farms, analyzing them and use results to anticipate decisions to be made by farmers, predicting farmers' demands to particular service and for a decision-maker to respond more effectively for future planning.

A decision made by a farmer in this study can be defined as an action or process of a farmer choosing or adopting one technology or services from the list of several alternatives. For the purpose of this study, three decisions were modeled include farmers decisions to choose certain breeding service (Either Artificial insemination or Natural bull), feeding animals concentrate (Farmer deciding to feed concentrate or not) and keeping of exotic animal (The number of exotic animals to be kept).

A farmer who is referred to in this study is a dairy farmer, farmers who keep cattle/cows for the purpose of producing milk. The study focused on small scale farmers, who most of the operations are based on low inputs, and small heard size. Depending on the countries, their herd size can range between 1 up to 20 cows. Also, production per dairy animal is low.

A decision-maker who is going to use the models developed in this study referred to all respective actors who guide, support, provide services and address farmers' needs to improve the quality of farming and ensure high productivity. The actors include policymakers, service providers, extension workers, researchers, investors, and other livestock stakeholders. Within this context, a policymaker is the one who creates plans of actions that farmers follow. While service providers can be a vendor or supplier who provides services or farm inputs to farmers. Therefore, the purpose of the study is the development of "what if" scenarios to be used by the decision-makers to anticipate preferences, demand, and the likelihood of technology, service,

program or investments to be adopted by farmers. Also, for the decision-makers to be able to identify forces that drive a particular livestock production system.

CHAPTER TWO

LITERATURE REVIEW

This study reviewed different publications on how ML has been employed in different dairy production systems. Specifically, the report aims to provide an overview of ML potentials in dairy. Congruently its practical challenges towards its adoption to small scale dairy systems were observed and provide recommendations focused upon the use of ML to support decision making to other stakeholders of livestock sectors such as policymakers.

The search for articles involved two databases: ScienceDirect and IEEE Xplore. Also, other articles were searched in Google Scholar which is a web scientific indexing service. The first step was to search for the articles. A searching query of keywords was created. The following queries were used “Machine learning” AND “Dairy” OR “Livestock”, another query was “Prediction models” AND “Dairy” OR “Livestock”.

The second step was to select all the papers that were relevant to this study. This study targeted any area of application in the dairy sector without limiting studies on disease control, policy, animal breeding, reproduction, production and farm management. Then, in the last step, all papers selected were reviewed. In total 44 papers were reviewed to analyze the problems that have been addressed, a solution proposed, ML algorithms used, and the nature of the study including study site. Also, other information gathered included the type of data used and if possible, data collection devices that were employed in the implementation of the study.

2.1 How machine learning is being used by other production systems

It has been argued that the adoption of advanced technology has a role to play for the success of a dairy farm (Awasthi *et al*., 2016). This theory has also been proven by large scale dairy farmers (Commercial farmers). The majority of large-scale farms have adopted advanced management technologies such as automatic milking systems, which are known to reduce farm input costs by reducing the number of workers on the farm (Heikkilä, Myyrä & Pietola, 2012). Developments of farm technologies have even led to fully automated systems that apply machine learning (ML) models for animal monitoring, disease detection and efficient use of farm resources.

Additionally, these techniques assist in monitoring livestock and livestock products through the value chain from farm to consumer (traceability) (AU-IBAR, 2015; Caporale, Giovannini, Francesco & Calistri, 2001; McKean, 2001). The use of technologies that utilize ML techniques is now becoming popular in developed countries especially in large commercial farms. However, the use of such technologies in small-scale farms is sparse.

Figure 4 shows how ML has been used in the dairy sector. Predominantly, information is collected from the farm using different electronic devices including; stationary observatories, animal mounted gadgets and hand-held tools empowered by different sensors. A global positioning system (GPS) and satellite are also being continuously used. In the end, ML comes in hand to extract/identifying patterns from collected data. Therefore, its main task is to convert raw data into valuable information that assists users to make informed decisions.

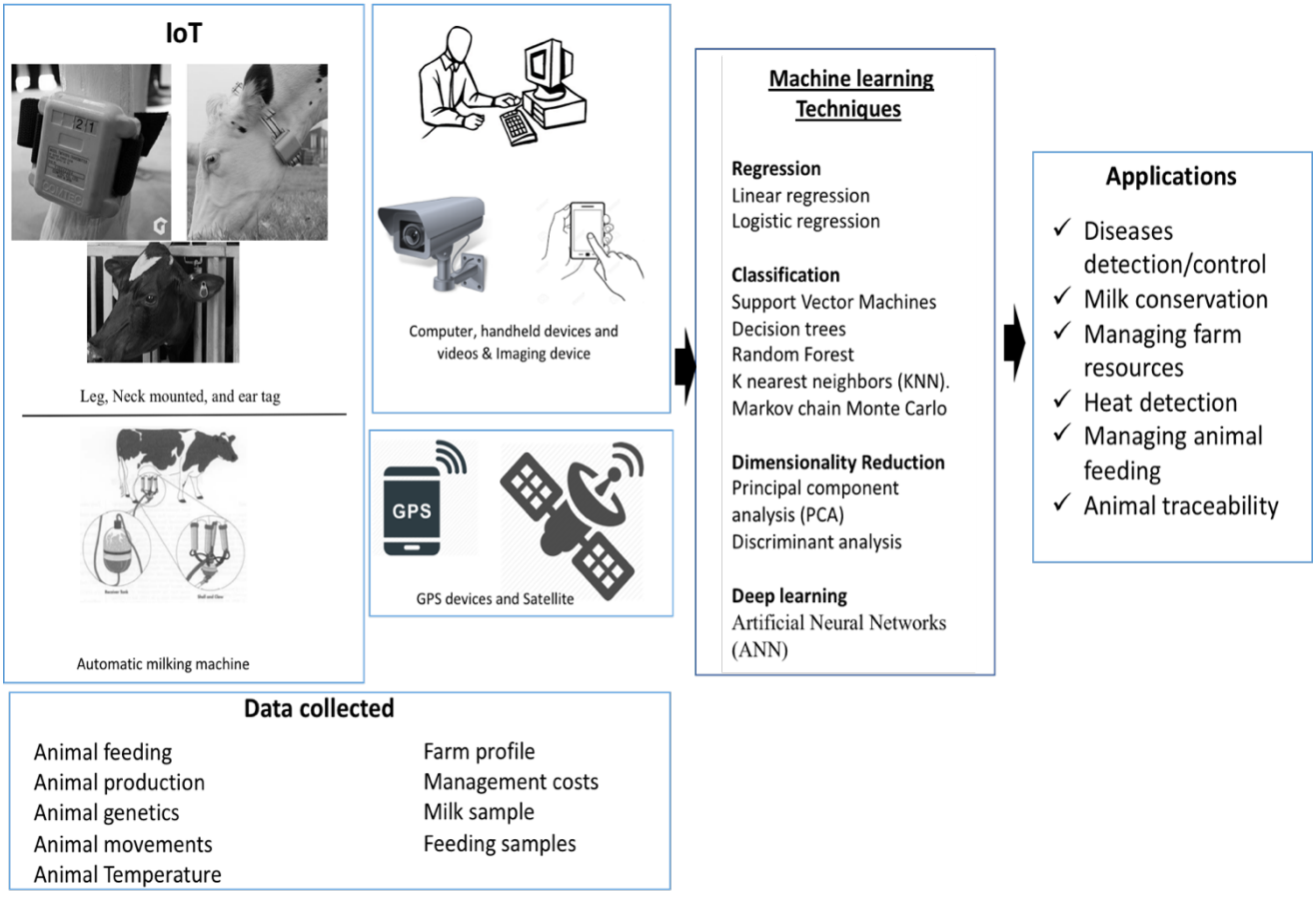


Figure 4: Summarize how various technologies and machine learning techniques has been used to add value in the dairy sector.

Table 1: Presents the list of papers that rereviewed in this survey. These articles were categorized based on their application; animal feeding (4 papers), animal replacement (1 paper), diseases detection/control (11 papers), farm management (1 paper), milk conservation (3 papers), reproduction and breeding (15 papers) and animal traceability (9 papers). From the total list of articles that were reviewed, 60% of these articles addressed the challenge of animal disease control, reproduction, and breeding. The table also shows different types of data sources that were used for the reviewed articles.

Table 1: Dairy areas where machine learning techniques were employed.

Area of Application	**Number of papers**	**Reference**	**Overview**
Animal feeding	4	Ali *et al*. (2014), Dórea *et al*. (2018), Roland *et al*. (2018), Chelotti *et al*. (2018)	These articles described how can ML techniques can be used in measuring and analysing feeding traits for an individual or group of animals in commercial dairy farms.
Animal replacement	1	Shahinfar *et al*. (2014)	The objective of this study was to investigate the potential of ML in order to estimate breeding values of a dairy cattle.
Diseases detection/control	11	Alsaaod *et al*. (2012), Goyache *et al*. (2005), Viazzi *et al*. (2013), Mammadova & Keskin (2013), Kamphuis et al. (2010), Barker *et al*. (2018), Zhao *et al*. (2018), Ebrahimie *et al*. (2018), Yazdanbakhsh, Zhou & Dick (2017), Amrine, White	These studies focused in establishing ML tools for early detection of diseases and assessing how can ML play an important role inreducing thenegative impact of livestock disease, increases the treatment success, and prevening the diseasess from becoming chronic.

		& Larson (2014), Parker Gaddis *et al*. (2016)	
Farm management	1	Shine *et al*. (2018)	The main goal for this study was to develop ML tool for predicting/forcasting farms costs i.e electricity and water consumption ondairy farms.
Milk conservation	3	Wei, Wang & Zhang (2013), Ma *et al*. (2018), Zhang *et al*. (2014)	These studies focused in developing smart system using ML techniques to be used in measuring the quality of milk i.ediscriminate adulterated milk from raw cow milk and in monitoring the quality and storage time of unsealed pasteurized milk.
Reproduction and Breeding	15	Caraviello *et al*. (2006), Keegan, Cunningham & Apperley (1995), Borchers *et al*. (2017), Schefers *et al*. (2010), K Hempstalk, McParland & Berry (2015), Shahinfar *et al*. (2012), Pietersma *et al*. 2003, Saleh *et al*. (2014), Shahriar *et al*. (2016), Fenlon *et al*. (2017), Caroline *et al*. (2017), Rutten *et al*. (2016), Dolecheck *et al.* (2015a), Cook & Green (2016), Borowska *et al*. (2018)	These articles tried to resolve different challenges faced by the farmers; where they focused in establishing ML tools that can predict calving in dairy cattle using information like, animals' behaviors. But also, other studies attempted to predict the likelihood of conception occurring and predicting conception outcome under different scenarios.

Traceability	9	Kumar *et al*. (2018), Hadad, Mahmoud & Mousa (2015), Dutta *et al*. (2015), Rahman *et al*. 2018, Smith *et al*. (2016), Williams, Mac Parthaláin, Brewer, James & Rose (2016), Benaissa *et al*. (2017), Nasirahmadi, Edwards & Sturm (2017), Santoni, Sensuse, Arymurthy & Fanany (2015)	The focus of these articles was to asses and develop ML tools to be used in identifying animal behavior, in which its pattern is linked to animal health and feeding e.i. To monitor animals behavious and compare the identified pattens to the list of classified animal behaviors and be able to detect if the animal is sick. But also identifying animlas preferences in regards to feeding.

2.1.1 Machine learning in enhancing animal reproduction and breeding

There is a number of applications for ML specific in animal reproduction and breeding. First ML is used to improve the accuracy of heat detection (estrous) (Keegan *et al*., 1995; Shahriar *et al*., 2016). The study done by Keegan *et al*. (1995), demonstrated how milk records and animal behavior during milking can be integrated into ML systems to detect estrous. When an animal is on heat it tends to significantly reduce milk production and abruptly compensate that at the following milking. Also, in some cases, cows that usually have a well-defined position in the milking order will present themselves for milking well out of that sequence during milking. These behaviors are clearly visible in the raw data for only around 5-10% of animals. However, after integrating this information with ML it outperforms the farmer's ability. For example, in a study by Shahriar *et al*. (2016), through the use of ML technology farmers is able to identify/detect the event with the accuracy of up to 82%. This was 32.7% more than a farmer would identify (Roelofs, López-Gatius, Hunter, van Eerdenburg & Hanzen, 2010).

Table 2: A summary table showing different farm data sources in a Dairy farm

Source of data	Application	Reference
Farm records: Use farm Databases (23)	Animal feeding	Dórea *et al*. (2018), Chelotti *et al*. (2018)
	Animal replacement	Shahinfar *et al*. (2014)
	Diseases detection/control	Ebrahimie *et al*. (2018), Alsaaod *et al*. (2012), Goyache *et al*. (2005), Mammadova & Keskin (2013), Parker Gaddis *et al*. (2016)
	Farm management	Shine *et al*. (2018)
	Reproduction and Breeding	Borchers *et al*. (2017), Cook & Green (2016), Borowska *et al*. (2018), Schefers *et al*. (2010), Caraviello *et al*. (2006), Caroline *et al*. (2017), Rutten *et al*. (2016), Dolecheck *et al*. (2015), Keegan *et al*. (1995), Hempstalk *et al*. (2015a), Fenlon *et al*. (2017), Shahinfar *et al*. (2012), Pietersma *et al*. (2003), Saleh *et al*. (2014)
Farm Samples (Milk) (2)	Milk conservation	Wei *et al*. (2013), Zhang *et al*. (2014)
GPS data (2)	Traceability	Williams *et al*. (2016)
	Diseases detection/control	Amrine *et al*. (2014)
Satellite data	Animal feeding	Ali *et al*. (2014)
Media (Images and Videos (6)	Diseases detection/control	Viazzi *et al*. (2013), Zhao *et al*. (2018)

	Traceability	Kumar *et al*. (2018), Hadad *et al*. (2015), Santoni *et al*. (2015), Nasirahmadi *et al*. (2017)
	Animal feeding	Roland *et al*. (2018)
Sensor data (Leg and Neck mounted collar mounted accelerometers Mobile sensor, ear tag (11)	Diseases detection/control	Kamphuis *et al*. (2010), Kamphuis *et al*. (2010), Barker *et al*. (2018), Yazdanbakhsh *et al*. (2017)
	Milk conservation	Ma *et al*. (2018)
	Reproduction and Breeding	Shahriar *et al*. (2016)
	Traceability	Smith *et al*. (2016), Rahman *et al*. (2018), Dutta *et al*. (2015), Benaissa *et al*. (2017)

Furthermore, electronic technologies are considered to be efficient than the traditional method, where a farmer has to visually observe the animal (Dolecheck *et al*., 2015). Visual observation is considered to be a very time-consuming where a farmer has to invest significant time in observing animals' multiple times per day (Shahriar *et al*., 2016). The process is also challenging to a free grazing farm and the exercise became even more complex in a large heard size. It is reported that most of the time farmers fail to detect or can detect while at a late stage of estrous (Saint-Dizier & Chastant-Maillard, 2018).

However, this previous model developed by Keegan *et al*. (1995), favored farms that use automatic milking systems. In this case, milk records were crucial in prediction, which can fail to work with non-automatic farms due to missing data when a farmer fails to record/observe. However, the gap was bridged by the latest model implemented by Shahriar *et al*. (2016), where biosensors (accelerometers) were used to collect data for predicting. This allows the auto collection of animal movements that were used in modeling instead of milk records. The use of accelerometers to collect data for heat detection was also suggested by review studies

conducted by Dolecheck *et al*. (2015) and Saint-Dizier and Chastant-Maillard (2018). Where they saw the need for implementing automatic data collection tools such as accelerometer in dairy farms which can be used to detect animals' behavior for pattern identifications. This technology is reported to enable a farmer to optimize herd reproductive performances and reduce the cost of hiring a number of farm laborers for monitoring a farm.

Also, ML techniques have been extended to ensure a successful conception rate. By developing models that can identify factors associated with successful conception rate (Schefers *et al*., 2010) or that hinder a successful consumption (Hempstalk *et al*., 2015a). Such models are important in measuring animal fertility and to identify factors that can hinder a successful conception. This information can be useful in decision support tools or even in selecting animals with good genetic merit. After identifying the factors that can hinder a successful conception rate, other studies focused on developing models to predict the insemination outcomes using production, reproduction, health, and genetic information (Hempstalk *et al*., 2015a; Rutten *et al*., 2016; Saleh *et al*., 2014). The study done by Cook and Green (2016) used information such as the amount of milk produced, fat and protein content to predict the likelihood of conception occurring by day 100 and 150 of lactation. This model can be used when the farm failed to produce a complete set of data required for prediction. Predicting the success of conception is essential to a dairy farmer as such information can guarantee a farmer in selecting/deciding future mating plan e.g. inseminate their animals with more expensive semen or choose to use low cost semen when the model predicts low a likelihood of conception (Hempstalk *et al*., 2015a; Hermans *et al*., 2017). This technology was also extended to accommodate pasture-based dairy farms as established by Caroline *et al*. (2017).

Animal calving is also one of the major events because of its crucial importance in herd economics and the amount of time required for its detection (Saint-Dizier & Chastant-Maillard, 2018). Providing timely calving assistance can reduce the risk of dystocia. Dystocia has severe consequences for the welfare of both the dam and calf, including pain, increased risk of surgery, the morbidity linked to other diseases, mortality, and culling (Fenlon *et al*., 2017). Therefore, it's always recommended to assist the animal during labor. However, approximating the time for labor is always a challenge to a farmer. Currently, dairy producers have been using the combination of breeding records and visual cues to estimate calving time; however, even experienced personnel may not accurately detect all calving, because of perceptible behavioral and physiological changes do not occur for every cow or at a consistent time across calving.

To date, the application of precision technologies (including ML) can use a combination of animals information such as animal activity, rumination time, and lying behavior to automatically detect and alert a farmer 8 hrs period before calving (Borchers *et al*., 2017; Fenlon *et al*., 2017).

Machine learning also revealed its importance to other companies providing dairy services including AI stations. The quality of semen used by a farmer has a significant economic trait in cattle as it determines the success of conception rate. Semen quality can be determined by a complex set of traits such as environmental factors, animal genetics, etc. Untying this complex interrelated relationship is difficult without using computation tactics such as ML. The study that was done by Borowska *et al*. (2018) showcased how ML techniques can achieve some of that. Borowska *et al.* (2018) investigated the genome regions associated with eleven semen quality variables of bulls, using ML analysis. Similarly, in another initiative ML was extended to compute the breeding values of Holstein cows (Shahinfar *et al*., 2012); which helped to predict the milk quality (Fat and protein content) to be produced. This technique has outperformed the old methods by inventing a rapid and low-cost solution. In contrast with the old methods that had a computational challenge and was time-consuming, where data to be used had to be recorded only periodically (e.g., quarterly or semiannually). While with Shahinfar *et al.* (2012) the method, animal performance data combined with breeding values of their parents is used. This information is easy to obtain on any farm. This allows Rapid identification of superior animals that can lead to earlier collection and distribution of semen and more rapid genetic progress.

2.1.2 Machine learning for diseases control

The most efficient way of managing a livestock disease is to detect and treat it before it either gets severe or spreads to other animals on a farm. For this reason, there has been a significant trend in considering the application of ML techniques in animal disease control (Yazdanbakhsh *et al*., 2017). This technology has gone beyond from only detection, to monitoring an animal during its treatment and afterward (Amrine *et al*., 2014). Farmers have been facing challenges and difficulties to detect animal diseases including lameness especially at the early stages of a disease. At first, detecting a disease depends on farmers' skills to visually observe animals; This method is time-consuming because farmers have to always be on a farm and often, they fail detected. The technology also was improved from detecting a single animal to a group of animals by using video recording data which is very essential to a big heard size (Alsaaod *et*

al., 2012). However, the use of accelerometers as the source of data has helped to improve the performance of the models (Zhao *et al*., 2018).

The opportunities for ML to minimize the accuracy of mastitis disease on farms is growing. Mastitis has been one of the common diseases on dairy farmers and causes a major loss to a farm after their milk being rejected by the market. The most commonly used methods are the use of somatic cell counts (SCC) and electrical conductivity. Relying on these methods alone raises a major concern for their reliability (Goyache *et al*., 2005). Because SCC lacks other key information e.g. Tremendous seasonal or age-dependent variation between SCCs. While electrical conductivity requires a very sensitive set of data to predict. This led to the creation of many false alarms and sometimes farmers experienced difficulties to define the threshold. But integrating these methods with ML has helped to improve the accuracy and reliability (Mammadova & Keskin, 2013; Parker Gaddis *et al*., 2016). This also assisted farmers to define the best threshold (cutoff) of different predictive milking parameters (Ebrahimie *et al*., 2018). ML models are also integrated with the robotics milking system where data that is collected from a robot is used instantly to test for diseases (Kamphuis *et al*., 2010).

Apart from ML depending on data from robotic milking, the use of other animal sensors, coupled with an intelligent surveillance system was also recommended as the best tool to quickly collect data from animals (Yazdanbakhsh *et al*., 2017). Sensors can collect data rapidly which can be used to predict on time and alert a farmer on time before contagious diseases spread; conventionally can save a herd from increased morbidity and mortality.

2.1.3 Machine learning for animal monitoring and traceability

Animal identification systems have taken a chart mostly in developed countries including Europe and America. The commonly used methods are handcrafted texture feature extraction and animal appearance-based feature representation techniques. These techniques are unable to perform animal recognition in the unconstrained environment (Santoni *et al*., 2015). Recently ML approaches have achieved more attention for recognition of species or individual animals using visual features e.g. muzzle point (nose pattern) (Kumar *et al*., 2018). This has helped in addressing the problem of missed or swapped animals and false insurance claims.

However, in dairy, the technology goes beyond that. Due to the increasing growth of the world trade and growing concerns of food safety by consumers, farmers are demanded to implement identification and traceability systems for their animals. This goes far to monitor an animal

throughout its lifetime. Which includes monitoring animal feeding, health and all other activities associated with the animal. Individual animal identification could be achieved by different methods, mechanical, electronic and biometric. The most commonly adopted technologies and which are popular all over the world include animal recording, ear tags, tattooing muzzle ink printing, and freeze branding and hot-iron branding. Others which are not commonly used include: Electronic Identification; Radio Frequency Identification (RFID) and DNA. Due to their weakness as an example: mechanical methods are not suitable for large-scale deployment. Also, can cause animal infections, and are not sufficient for traceability purposes. With that regard, scientists have tried to minimize the risks by using ML models to uniquely identify animals. Instead, muzzle print, which is similar to the human's fingerprint, has proven to be a unique feature of each cattle (Tharwat, Gaber & Hassanien, 2014). Therefore, ML algorithms come on hand to uniquely classify muzzle prints. This also has helped to maintain the safety of animals suffering from diseases such as bovines and during guarantees of the livestock products (Hadad, Mahmoud & Mousa, 2015).

The continued use of ML linked to wearable animals' sensors offers a farmer the potential for continuous and autonomous monitoring of cattle without the need for human involvement. The techniques can be employed for many functions including animal behavioral classification (Benaissa *et al*., 2017; Dutta *et al*., 2015; Smith *et al*., 2016). ML has been used to identify animal behavior, in which its pattern is linked to animal health to compare the classified behaviors to rules regarding the animal's expected or normal behavior hence enabling early detection of animal sickness. Also, it can consistently allow monitoring of feed intake and safety of animals (Rahman *et al*., 2018) which is more efficient and become more useful to an intensive grazing farming system (Rahman *et al*., 2018). Because at some point a farmer needs to know the amount of feed intake compared to the amount of pasture or supplements offered or animal preference. Therefore, monitoring of animals' behavior has become important (Nasirahmadi *et al*., 2017).

2.1.4 Machine learning on animal traceability and feeding

Animal traceability goes beyond monitoring of an animal throughout its lifetime and this includes monitoring what they feed. Knowing ***how*** the animal feed is of benefit to a farmer. But knowing ***what*** they feed is one of the key information that must be known by the entire value chain including consumers and food processors (AU-IBAR, 2015; Caporale *et al*., 2001; McKean, 2001) and this information should be traceable (Caporale *et al*., 2001; ISO, 1995).

GPS is a satellite-based radio navigation system that provides geolocation and time information to a GPS receiver anywhere on or near the Earth. It is widely used in different fields to collect information for supporting decision tools. It has also gained its popularity based on its functionality including; First is a free data collection tool, it just sends information without requiring the user to transmit any data, it operates independently of any telephonic or internet reception. That makes it cheaper and it does not add additional costs to users for using the internet to connect. In dairy, it is now commonly used in monitoring animals, especially in a pasture-based dairy practitioner during grazing. Therefore, animals will need GPS collars to connect with GPS for monitoring during feeding. This technology can track the animal and provide information such as animal behaviors during grazing, resting, and walking. Then ML comes on hand to search for patterns in data that are unobservable by the human eye (Williams *et al*., 2016).

ML can be intergraded with GPS to assist farmers in estimating the biomass in intensively managed grassland farms using vegetation indices data (Ali *et al*., 2014). One of the weaknesses of GPS is it can be blocked by environmental setting e.g. building. Therefore, zero-grazing farms still can use other methods e.g. biosensors to monitor animal behavior during feeding. Animal sensors can record data about animals such as bite/chew behavior and link to ML for pattern identification (Chelotti *et al*., 2018). This initiative was also extended by Dórea *et al*. (2018) where instead of animal behavior, milk samples were collected from the animal, then used to predict the amount of dry matter to be consumed by the animal. Also, an accelerometer which is considered to be more reliable for zero-grazing dairy farms was used to collect animal drinking behavior for dairy calves (Roland *et al*., 2018). This information is essential in monitoring dairy calves where a farmer can be prompted in case of any changes for interventions. This assists a farmer in reducing the negative effect on calves' health and weight gain.

2.2 How can machine learning help to improve small scale farmers' productivity

From the studies reviewed, it's obvious that the use of advanced tools such as ML and other animal tracking devices has a significant role to play in maximizing farm production and minimizing running costs. Our study has demonstrated that ML has a potential role to play in addressing some of the challenges in the dairy sector. Since small-scale farmers do not spend full time on their farms because they are predominantly mixed farms. Technologies can be of great value to them. Machine learning has the potential of changing the old ways of doing

farming. Instead of farmers being present physically all the time to monitor their farms, with these advanced technologies they can monitor their farms from a distance.

Inline with the preceded paragraph, it was reported that up to now farmers are still using a conventional method for tracking their animals which are not suitable for proper identification/traceability (Grace, 2013). As a result, many farmers fail to meet the rules set by the World Trade Organization (WTO) (Thiermann, 2005; Zepeda *et al*., 2005). For farmers to be able to meaningfully gain access to international markets they need to abide by international standards i.e. international diseases control standards and animal traceability (Brester, Marsh & Plain 2003; Gimeno, 2003). Therefore, it's time for farmers to start considering the use of these technologies (ML) on their farms.

Heat detection has been one of the challenges facing small-scale farmers. This problem is contributed by a lack of supporting tools for heat detection. In the study done by Mwanga *et al*. (2018) it was indicated that most of the small-scale farmers preferred visual observation as the main method for heat detection and ensuring timely insemination of their cows. Moreover, it was also reported that mostly don't keep records (Dudafa, 2013) while other farmers fail to produce any records. Those who try to keep data use very poor recording systems (Brooks-pollock *et al*., 2015). Books, papers, writing on walls have been the main tools farmers use to keep records while others rely on their memory (Chagunda *et al*., 2006). These methods have proven failure as it was indicated that using animal tracking systems and ML techniques were more consistent and more accurate than relying on human observations. It has been reported that for every missed estrous a farmer will incur (8%) loss of total milk production which is equivalent to a cost of 21 days without milk (Mitchell, Sherlock & Smith, 1996). This is a common situation faced in many small-scale dairy farms. Therefore, in order to continue stabilizing small scale farmers economy, there is a need for adopting advanced technologies as it was proposed in this study.

Most small-scale farms depend on extension officers. Extension officers need to assist farmers in their day to day activities. But in developing countries, most farmers do not receive the service due to the limited number of extension workers. Therefore, it is time for small scale farmers to opt using advanced technology which can extend extension service to their farms by giving informed decisions on time. As it was specified in this study that farmers would be able to know various animals' behaviors from data that is extracted from advanced technology. However, it is recommended that the implementation of these smart farming

systems should reflect the context of a small-scale farmer. For example, the use of local language and voice user interface (VUI) with concise and simple messages.

Based on the challenges presented above, the use of sensors to collect data can be one of the alternative solutions to small scale dairy farmers in bridging the gap. Therefore, data analytics i.e. ML would be needed to translate the measured sensor data into specific information e.g. animals' behaviors, detecting estrous or diseases. Sensors are believed to collect a large amount of information in a given time based on users' settings (Djedouboum *et al*., 2018). Though, one of the drawbacks of this technology, if have to be adopted by small scale farmers, is the installation costs. It is approximated that a single biosensor can cost up to $10/ per animal (SEMTECH, 2017). This obstacle can be dealt with and the possibility of having other supporting plans where subsidizing of these devices should be considered.

The adoption of these technologies to small scale farmers would still be challenged as there are a number of factors that need to be considered. One is the availability of data storage at the farm level. A number of small-scale farmers cannot afford data storage devices such as servers or computers. The use of shared servers can be an alternative where farm data can be stored to a public region or district server or data collection devices (animals' sensors) and embedded with analytics engine (ML) to process the data as is being collected.

Another challenge is the lack of reliable internet connectivity. Most of the small-scale farmers are located in rural areas where access to the internet is a problem. Hence provision of the internet to this farm will be of great value. But an alternative could be the use of cellular technologies. While there are efforts of connecting the world as one village, to the moment cellular networks can be an alternative. By leveraging existing infrastructure and mature technology data collection devices (sensors) can be integrated with existing cellular technologies which will allow millions of devices to be connected with little additional investment (Eric Conn, 2018). This can solve the identified obstacles. However, there will be still a need to integrate the databases with computation analytical programs (ML) and allow a farmer to receive only short reports via their mobile phones.

Other factors that will need to be considered is the access to complementary inputs, such as electricity. Electricity is reported to be a hindering factor in achieving smart farming. Electricity to a farm is not only essential to run machines or devices such as refrigerators. But

act as a supporting component to other electronic devices e.g. charging of mobile phones, computers, etc.

2.3 The use of machine learning to facilitate decision making by other livestock stakeholders

Nevertheless, it was interesting to note that there was limited representation of studies investigating the use of ML in enhancing other decision-making processes e.g. policymaking. This signifies that the use of ML algorithms to support policy decision-making in the livestock sector is in its infancy. As it was demonstrated in this study that ML has the potential in improving productivity, also there is a potential for the technology to do much more, in enhancing policymakers and other stakeholders in decision making.

Hence the same initiatives need to be extended to other livestock stakeholders. Designing policies and setting of strategies require appropriate planning of resources, prioritization of strategies, identifying farmers' demands and preferences (Hansson & Lagerkvist, 2016). The process is continuous. Hence the use of machine learning became even vital to other decision-makers such as a policy department.

It has been reported that a lack of decision support tools has cost the African continent especially developing countries. Where it has been difficult even for policymakers to extract valuable information from the farm (Dudafa, 2013). As a result, many plans, or strategies implemented often failed. Many programs and projects are poorly designed and inadequately targeted due to lack of decision supporting system which lead to the inefficient and fragmented allocation of scarce development resources (Ugo Pica-Ciamarra *et al*., 2014). Similarly, policies that are being designed are often incoherent with ill-defined goals and with little or no assessment of their likely impact. These have remains to be a major constraint to livestock development.

Moreover, planning, budgeting, and forecasting of the livestock sector are only a representative few of the many decisions facing policymakers (Pica-Ciamarra, Otte & Martini, 2010). To ensure that they meet farmers needs it requires them to trading off and makes decisions in regards to size, timing, investment of capital and prioritization of strategies (ILRI, 2018) while complying with farmers' preferences and demands (Tongeren, 2008). Similarly, the investors need to inquire about the location of operating units (where to invest), what to invest as well

as the timing of new investments (Tongeren, 2008). In regard to these challenges facing decision-makers every day, one can appreciate the value and impact of having decision-supporting tools to enhance evidence-based decisions.

Automation of decision supporting tools using ML in today's business processes has gone beyond the assembly lines of the past. It is now used in validating real-time business decisions such as market forecasting (Tongeren, 2008). Likewise, there are several complexities to each marketing decisions to be made. One has to know and understand customer needs and desires, while, having a good grasp of changing consumer behavior and align products to customers' needs and desires (Yeomans, 2015). Through Decision Support Systems, managers have been able to get reliable insight, predict consumer behavior. Even in recommending products to customers, it enables marketers to learn a user's content preferences and push content that fits those preferences (Rodrigues & Ferreira, 2016). Moreover, retailers now can accurately predict and respond to product demand and know more about how their products are received by their target audience. Comparing these to manual mining which requires long hours, ML has helped shorten this through reliable search and analysis functions. It has assisted decision-makers to quickly make decisions and take actions.

Joseph Byrum (2018) stated that agriculture has lagged on the data side because unlike the profitable market for consumer goods agriculture is often seen as a low-margin business with little opportunity for high-tech investment. Forgetting that these technologies are the one that is needed to drive productivity to the next level. As it was described above that there is a need of having decision supporting tools for effective policymaking and planning.

CHAPTER THREE

MATERIALS AND METHODS

In this chapter, we describe the research methods and strategies deployed to address our research questions and pursue our objectives. Figure 5 summarizes all the steps taken. The study started with data collection, which was collected electronically and stored on online servers. Then preprocessing of data including data cleaning and transformation of variables was performed. The modeling process started by screening features to be used in model development. Features obtained were used in the next step of model development. Finally, the algorithms used during model development were tested with a new set of data.

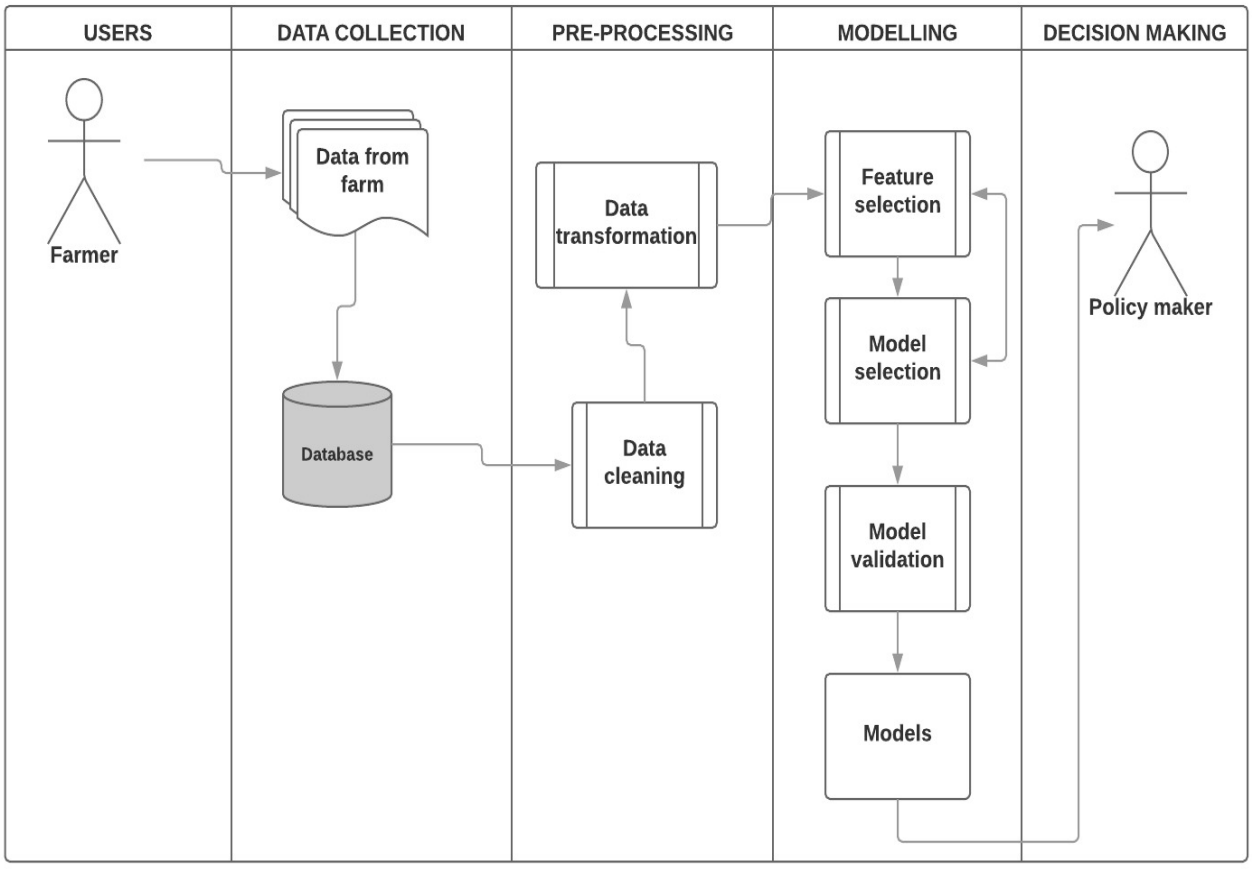


Figure 5: Conceptual framework for the study design employed in this study

3.1 Study sites

Data were collected in four countries: Ethiopia, Kenya, Tanzania, and Uganda. The study focused on dairy farmers; hence, study sites in traditional and emerging dairying zones were selected to maximize the number of dairy farmers to be included in the study. Figure 6 shows the study sites in the four surveyed countries. Four milk sheds were identified in Ethiopia: Addis Ababa, Asela, Bahir Dar, and Hawassa. In Kenya, three adjacent dairying zones were selected: Central, North Rift, and South Rift. In Uganda, three zones were selected based on their concentration of dairy activities. These were: Kiruhura, Wakiso, and Mbarara. In Tanzania, six regions were selected: Arusha, Kilimanjaro, Tanga, Iringa, Njombe, and Mbeya.

A cross-sectional survey was conducted through face to face interviews of the target farmers on their households over a one-year period (from June 2015 to June 2016). The list of oral questions used during the field study is attached to the appendix 1 structured questionnaire coded in Open Data Kit (ODK) was used to capture data electronically. A total of 16 308 small-scale dairy farmers were interviewed as follows: Ethiopia (4679), Kenya (5278), Tanzania (3500) and Uganda (2851).

The development of models depends on data that are of high quality and sufficiently. Therefore, this study employs various criteria for controlling data quality. Information such as the day, time of completion of a questionnaire and the geographical location (GPS) was implemented to the questionnaires to be used for quality control (QC). Incomplete and inaccurate questionnaires were also discarded. After data cleaning and quality checking processes, only 13 095 respondents qualified for inclusion in the analysis and these were distributed as follows: Ethiopia: 2892, Kenya: 4400, Tanzania: 3236 and Uganda: 2555.

3.2 Data and definition of variables

Selection of the factors that influence farmers decision was based on the domain knowledge and empirical findings from the literature. In analyzing the variables to be tested this study adopted the framework for the household system (FHS) (FAO 1990). This study investigated various factors that can influence household decisions and their interactions. This comprises of the following components: (a) farm management decisions which include factors such as farm investments, marketing, production and conservation decisions, (b) farm factors that included all farm

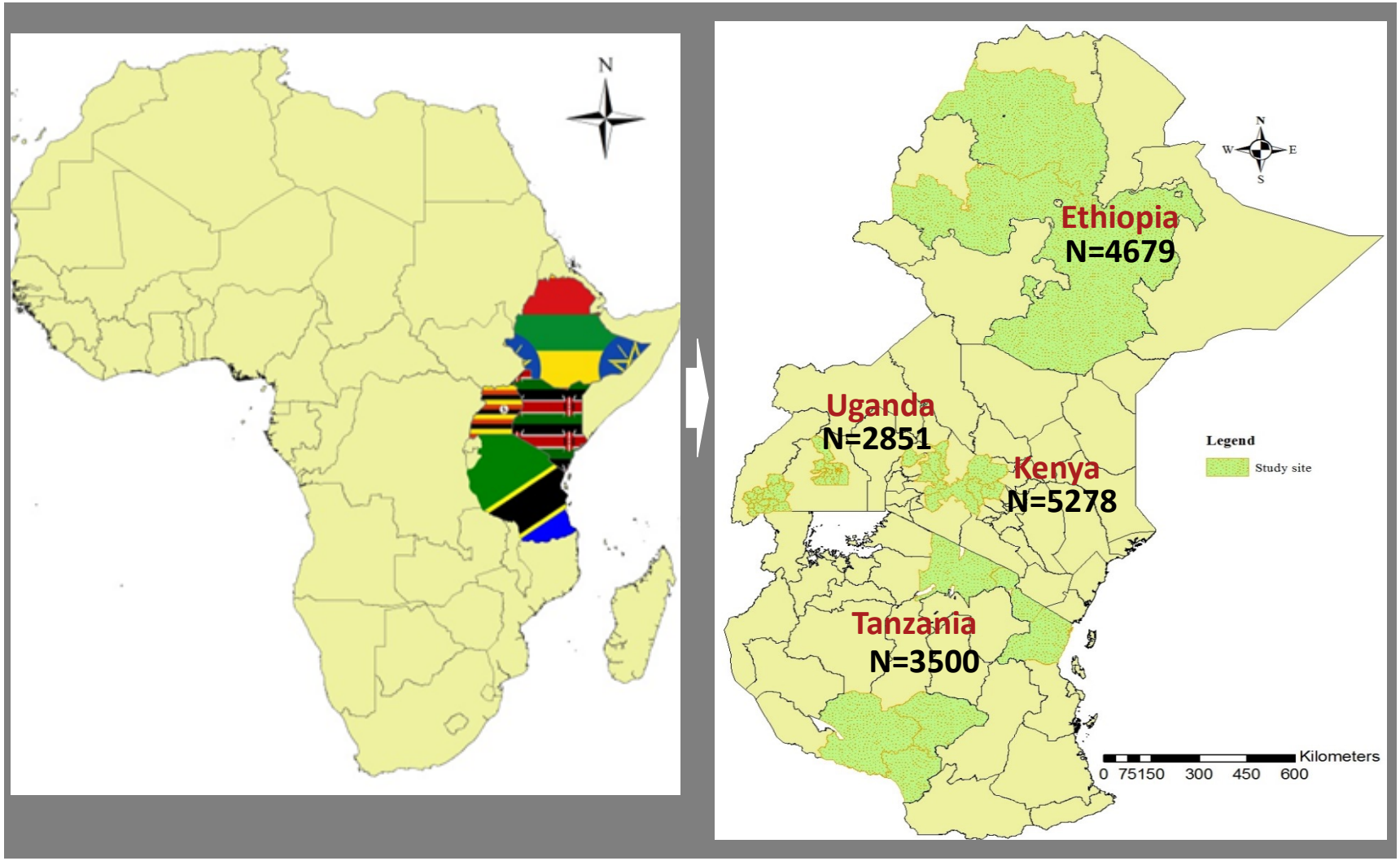


Figure 6: Map of the study regions; Ethiopia, Kenya, Tanzania and Uganda. The study focused on dairy farmers; hence, selection of study sites was done in traditional and emerging dairying zones

characteristics and can be broken down into socio-economic and biophysical factors. It includes information about climatically and geographical location, (c) off-farm factors that comprise a diverse set of factors: markets and market channels, policies, rules and regulations, support services and technical information. The proposed factors were also included as hypothesized features in developing models for this study. Figure 7 shows some of the factors that were tested grouped into four components as explained below. A complete list of all the variables that were tested is described in Table 18.

3.2.1 Farm characteristic variables

The following farm characteristic variables were analyzed: farm assets (land and herd size), a total number of laborers in the household and number of months a farmer purchased fodder, concentrate and crop residues in the year preceding the survey. Data on animal production was limited to estimated milk yield at the start of lactation, peak, and end of lactation for the best and worst animal in a herd. The values reported were average estimates based on farmer recall and not the actual realized yields. Two-factor scores were obtained after performing factor analysis on animal production variables including production at peak and lactation length for the best and worst animal. A score table (Table 3) was constructed by assigning different weights to several qualitative variables including binomial and other categorical measures based on their quantitative score.

3.2.2 Farmer characteristics

Evaluation of biographical variables included years of formal education and the total number of children in the household. Dairy management variables included: records keeping (1=Yes, 0=No) and methods for oestrus detection. Other variables included membership to a farmer group (1=Yes, 0=No) and household experience in dairy farming (0 for no experience;1 for one to five years; 2 for six to ten years; 3 for eleven to fifteen years; 4 for sixteen to twenty; 5 for twenty-one to twenty-five; 6 for twenty-six to thirty; 7 for thirty to forty; 8 for forty-one to fifty; 9 for more than fifty).

3.2.3 Infrastructural and institutional settings

Institutional settings variables included the following: number of times visited by an extension officer, distance to market in kilometers, availability of vaccination services (1=Yes, 0=No), availability of breeding services (1=Yes, 0=No), cost of breeding, distance to the service

provider in kilometers, availability of water (1=yes, 0=no), distance to market (kilometers), and transportation costs.

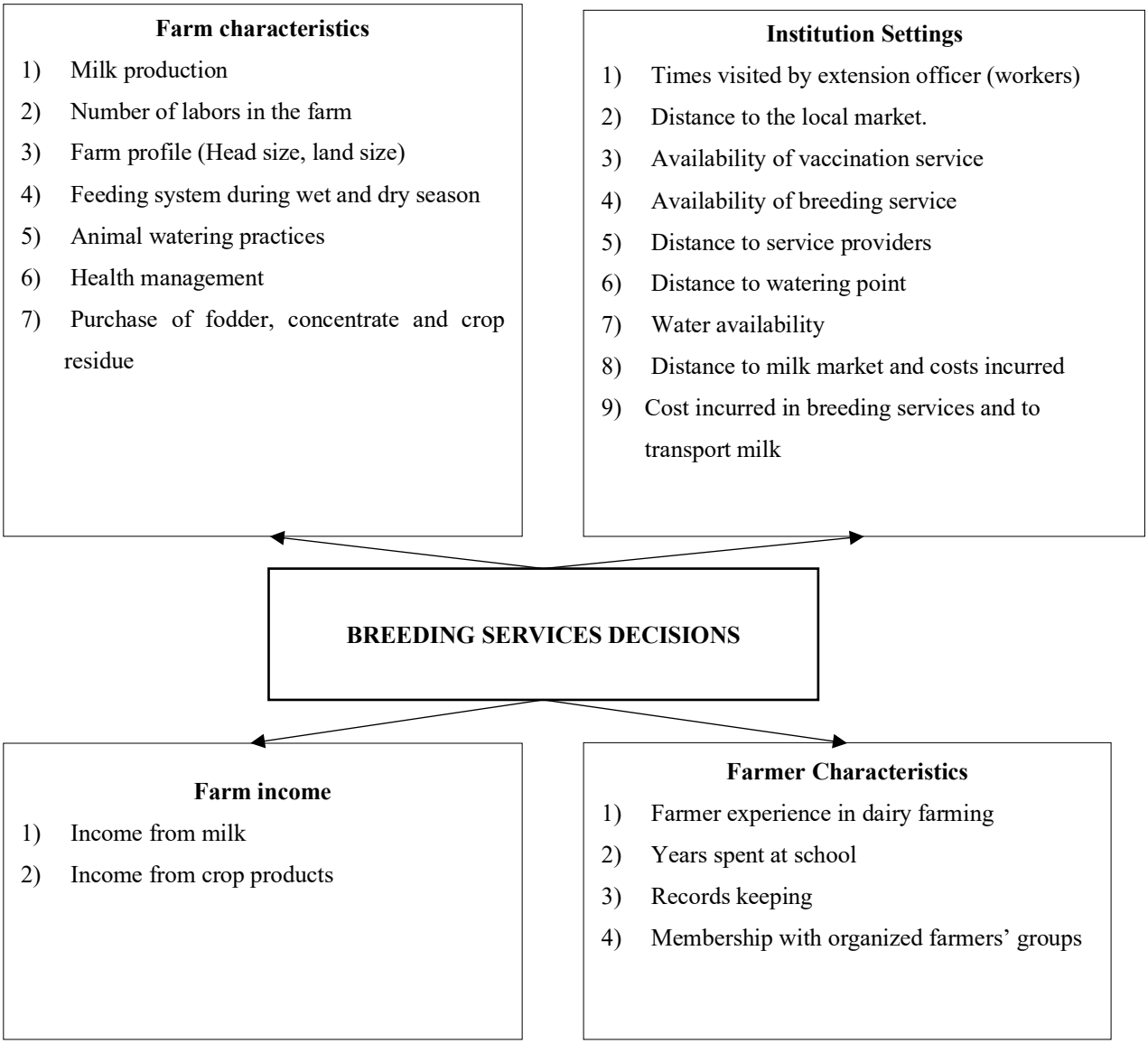


Figure 7: Conceptual frameworks highlighting various factor that can influence farmers decision. Four main characteristics of each business were included: Farm characteristics, Institutional settings, Farm income and Farmer characteristics

Table 3: Weight allocations for categorical variables

1. Binomial variables Yes=1, No=0 **2. Breeding methods** Bull=0, AI=1	**3. Household experience in dairy farming** No experience=0 one to five years=1 Six to ten years =2 Eleven to fifteen years=3 Sixteen to twenty years =4 Twenty-one to twenty-five=5 Twenty-six to thirty=6 Thirty to forty=7 Forty-one to fifty =8 More than fifty=9	
4. Distance to market One kilometer =1 Two kilometers =2 Three kilometers =3 Four kilometers =4 Five kilometers =5 Six to seven kilometers =6.5 Eight kilometers =8 More than eight kilometers =9	**5. Feeding System** Mainly grazing =1 Mainly stall feeding =2 Only grazing =3 Only stall feeding =4 Transhumance all animals =5 Transhumance some animals =6	**6. Frequency of treating cattle diseases** Never =0 Weekly =1 Two weekly =2 Monthly =3 Two monthly =4 Four monthly=5 Twice year =6 More than a year =7
7. Dewormingtimes never=0 once =1 Twice =2 Three times =3 Four times =4 More than four =5	**8. Times vaccinated** once =1 twice =2 Three times =3 Multiple =4 Other frequency=5	**9. Watering frequency** Do not water animals =0 Once=1 Twice=2; Thrice =3; Watering *ad libitum* =4;

3.2.4 Farm income

Two variables, income from crop sales and income generated from milk sales, were included to represent sources of farm income.

3.3 Methodology used for the first objective (characterize farmers decisions)

The first objective was to characterize farmers to develop an understanding of science domain but also to explore on different patterns of factors that can influence their decisions and determine if these factors are similar among regions. To achieve this, one decision was selected; farmers' decision to select a particular breeding method.

A multivariate logistic model was employed to explore factors that influence breeding decisions (Bull or AI). Two approaches were used in the data analysis. Figure 8 summarises the methodology that was employed to execute this objective. In the first approach, the t-test and Chi-Square test were used to evaluate whether there was a significant difference in the selected variables between farmers who use AI and those who use bull service. In the second approach, selected variables (and associated factor scores) that were hypothesized to influence farmers breeding choices were tested using a logistic regression model as follows:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e_1 \quad (1)$$

Where yi is a vector of the breeding method adopted by each farmer; β1, β2, β3, and β4 are vectors of coefficients associated with each explanatory variable category, x1, x2, x3, and x4 are incidence matrices that link the fixed effects of the explanatory variable categories (farmer, farm characteristics, income and institutional respectively), to the response variable; and e1 is the error term. All analyses were executed using SAS version 9.4 (SAS, 2003).

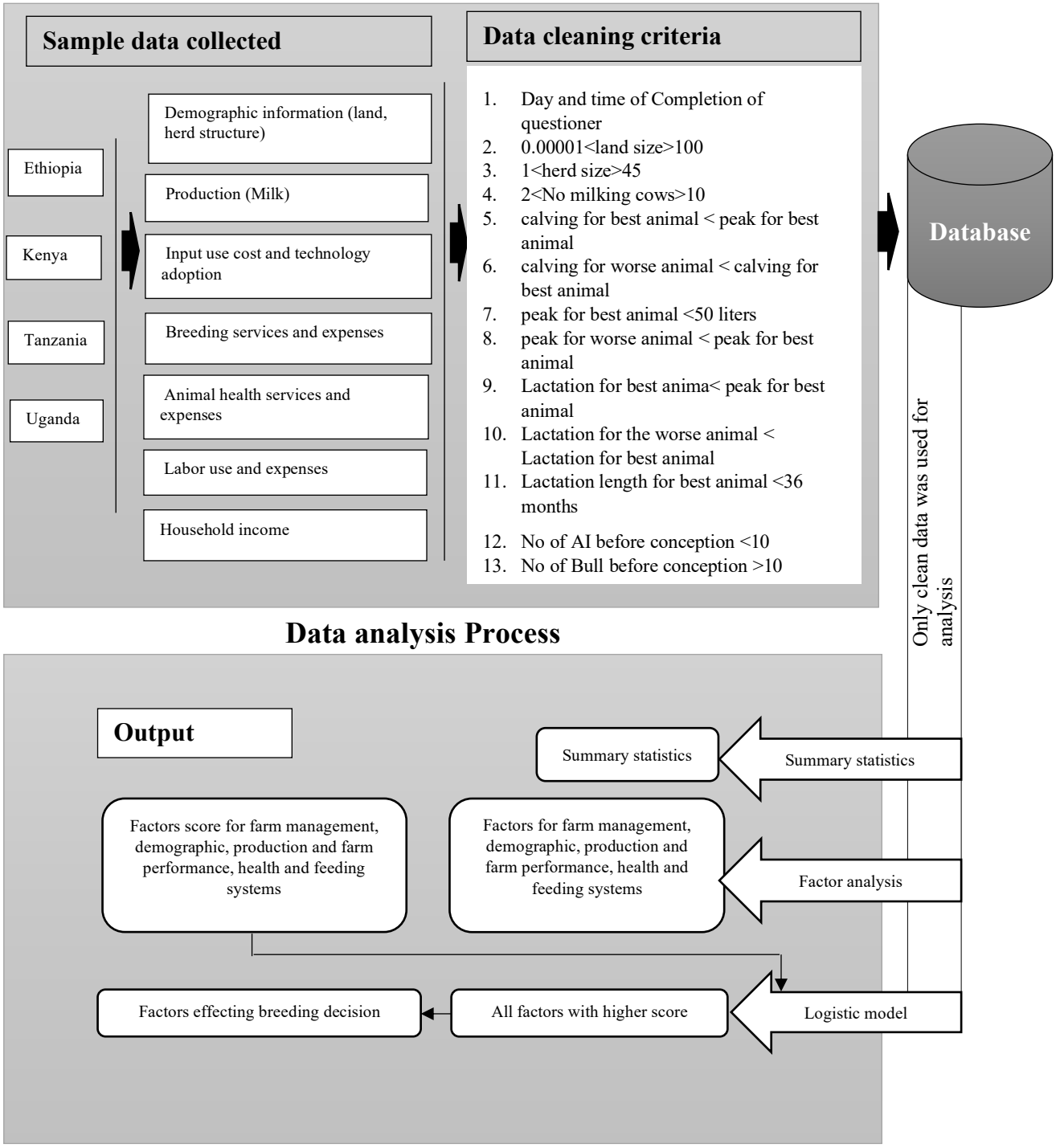


Figure 8: Summaries methodologies that were used for the first objective; To Characterize farmers decision making process in order to identify patterns of information that farmers use in making decisions

3.4 Methodology used for the second objective: Models development

The study modeled three decisions made by farmers. Three of them address farmer's decisions to use supplements, breeding methods and the number of exotic animals to be kept by a farmer. The study also investigated the key drivers that promote or hinder animal productivity and developed a model that predicts the amount of milk to be produced by an animal. In modeling animal productivity, the study refers to the productivity of the best animal during peak production.

The process of developing predictive models involved major two tasks: features selection/filtering and model selection. Figure 9 summarizes how the experimental study was carried out. Including data sources, machine learning methods, techniques used for features selection, model development and validation. The development of each model begins with features/predictors engineering. Three models were used for feature selections including random forest, Boruta, linear model for continuous variable and logistic for categorical variables. The top 15 features were selected for modeling. In predicting farmers decisions, six models were tested including linear regression for continuous variables or logistic regression for categorical variables, decision tree (DT), random forest (RF), K-nearest neighbor (KNN), Gaussian Mixture Model (GMM) and Artificial Neural network (ANN). A K-fold method was used for the model's validation.

Therefore, in this multistep modeling procedure, an outer and inner cross-validation loop was implemented. Meaning cross-validation was applied to the entire sequence of modeling steps (Friedman, Hastie & Tibshirani, 2001). Where samples were left out before any selection or filtering steps were applied as portrayed in Fig. 9. A ratio of 70% for training and 30% for testing was maintained. All analyses were executed using R software, version 3.5 running on a 64-bit system with 16 GB RAM and 2.70 GHz CPU. Table 19, shows sample R code implemented to develop models.

3.4.1 Data processing and variable selection

Considering that a dairy farming system is a result of a complex interaction of numerous interdependent mechanisms. This study adopted all variables that were hypothesized to influence farmers' decisions. In total, more than fifty features/variables were considered.

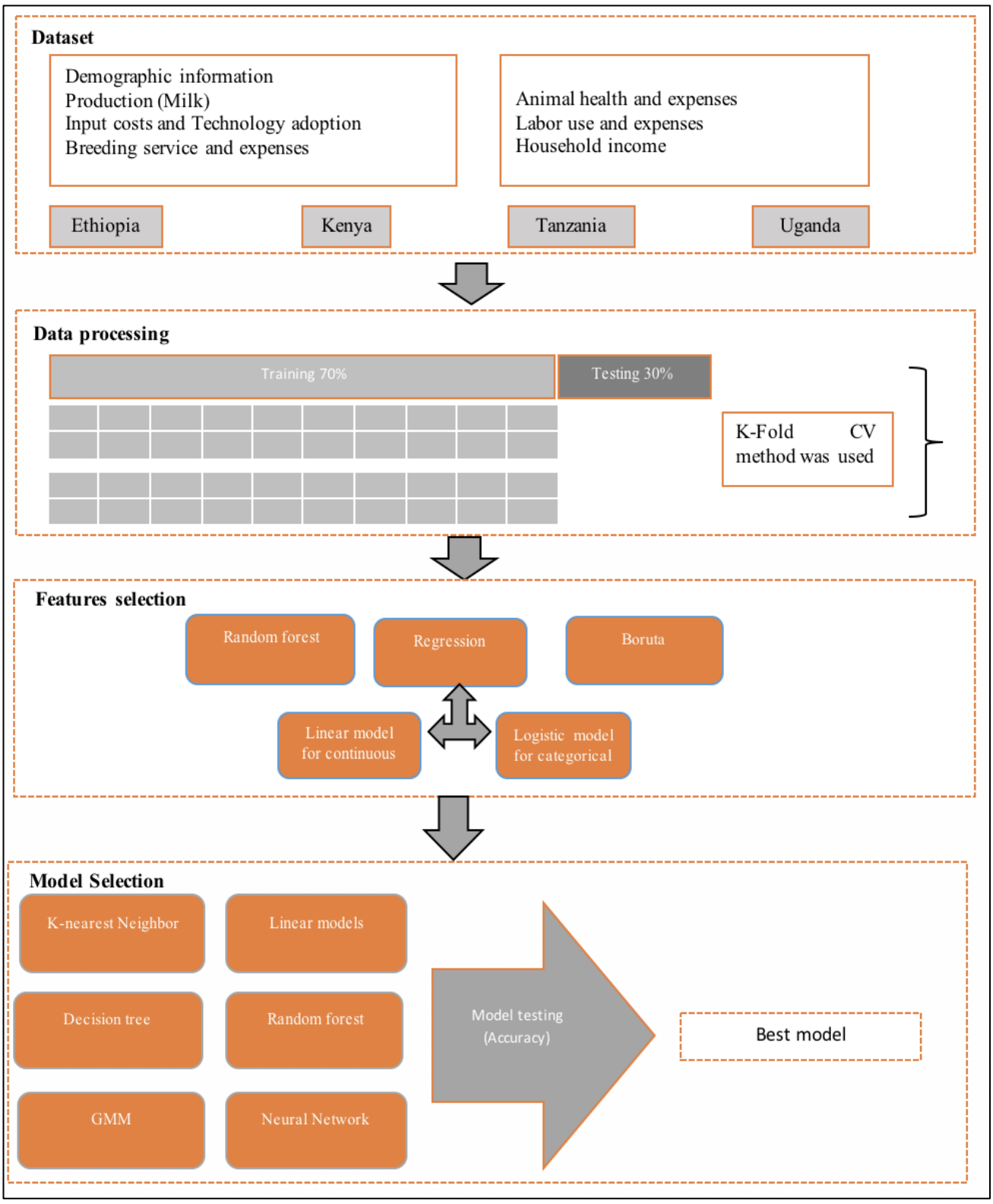


Figure 9: Machine learning framework employed in features selection, model development and validation

Also, in this step more, data processing was performed. All categorical variables were converted into dummy variables of 0 and 1 and make a total of more than 120 features.

This study also defined a range of the two continuous dependent variables based on the country profile: Number of exotic animals and amount of milk to be produced. Farmers who exceeded the specified limiting range were considered as outliers. Two approaches for identifying outliers were used; visual observation and DBSCAN. Visual observation was achieved through visualizing a frequency distribution of data into a histogram chart.

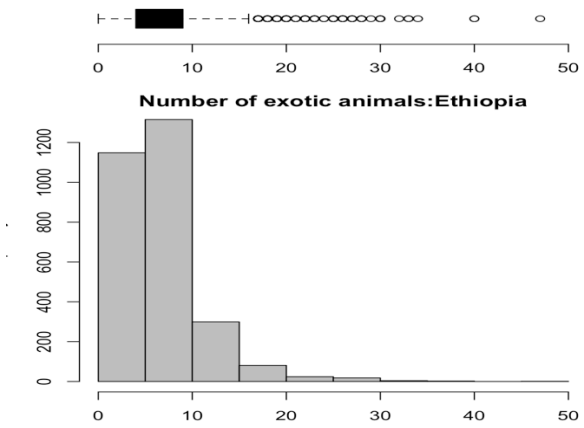
Figure 10 (a-d) shows a histogram distribution of the number of exotic animals respective for each country. Hence, farmers who were considered as outliers are farmers with > 20 animals in Ethiopia, >30 animals in Kenya, > 20 animals in Tanzania and >45 animals in Uganda. Animal milk productivity was taken at its peak. Farmers with less than 30 liters/day /animal were considered for all the countries except for Uganda where we included only farmers with less than 25 liters/day /animal as shown in Fig. 11 (a-d). Farmers with zero number of exotic animals were included with the assumption that a farm has no exotic animal neither produce milk.

3.4.2 Machine learning models

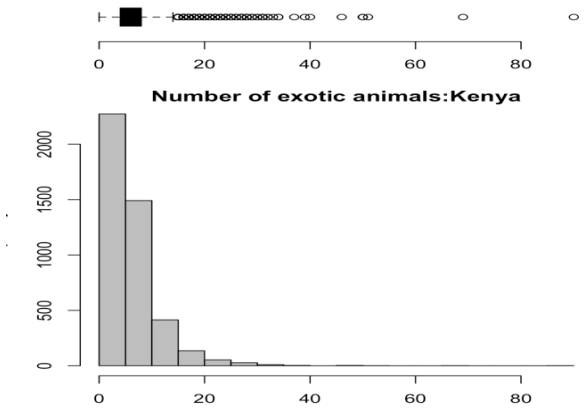
Three algorithms were used for features selection including Random Forest (RF), Boruta, linear model (LM) for continuous variables and Logistic Regression (LR) for categorical variables. The sets of features identified were tested in the four supervised ML models including; linear or logistic models, decision tree, Neural network, random forest, Gaussian Mix-Model and K-nearest neighbor.

(i) Linear and Logistic regression model

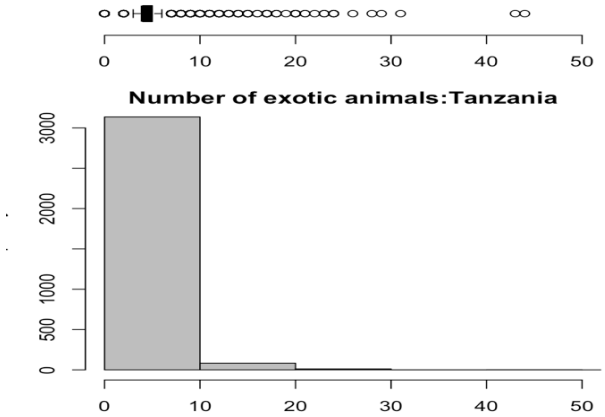
Linear regression is a common statistical technique used to express a class variable as a linear combination of the features. It was designed to predict real numeric value (linear models) but it was later modified to predict class values (logistic regression). Therefore, this study employed a linear model to predict numeric variables and logistic model for classification



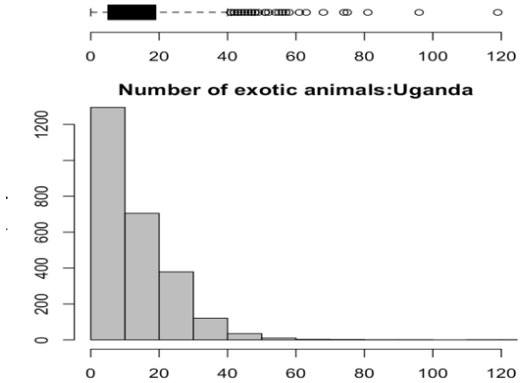
(a) Ethiopia: Farmers with more than 20 animals were considered as outliers



(b) Kenya: Farmers with more than 30 animals were considered as outliers

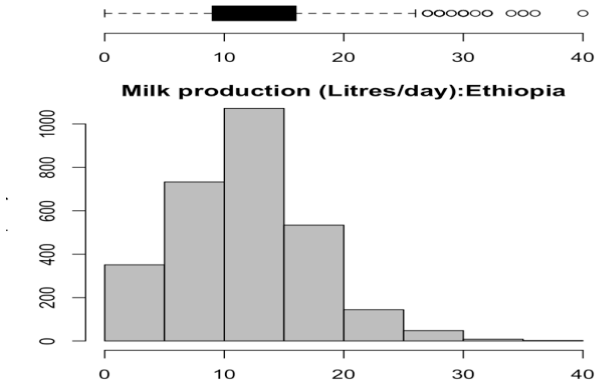


(c). Tanzania: Farmers with more than 20 animals were considered as outliers



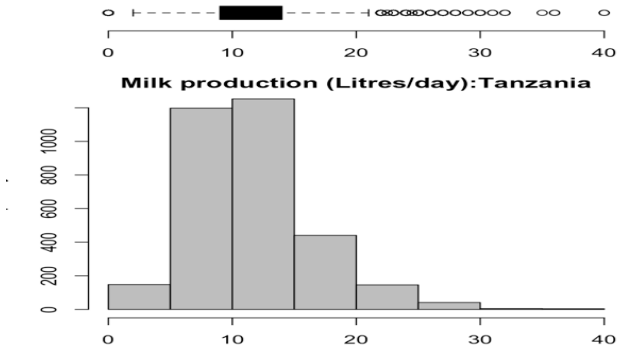
(d). Uganda: Farmers with more than 45 animals were considered as outliers

Figure 10:Histogram distribution of number of exotic animals respective for each country

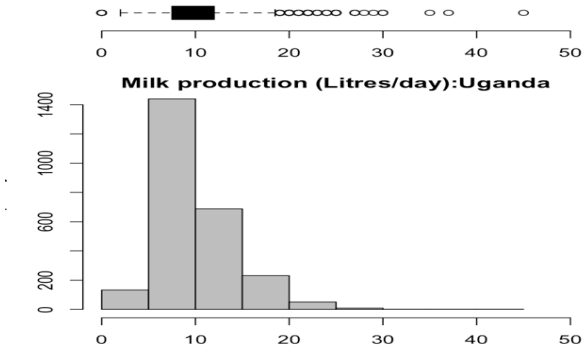


(a). Ethiopia: Farms that produce more than 30 liters/day were considered as outliers

(b). Kenya: Farms that produce more than 30 liters/day were considered as outliers



(c). Tanzania: Farms that produce more than 30 liters/day were considered as outliers



(d). Uganda: Farms that produce more than 25 liters/day were considered as outliers

Figure 11: Frequency distribution showing the amount of milk produced by the best animal/day

One of its assumptions is that there is one smooth linear decision boundary. Hence it works well when there is a linear relationship between the variables. In the case with nonlinear decision boundary, a probabilistic assumption is preferred where predictions are mapped to be between 0 and 1 through the logistic function P(Y|X). The logistic function works well when the classes are linearly separable (i.e. can be separated by a single decision surface). Linear models are intrinsically simple, have low variance and so are less prone to over-fitting due to the regularization techniques. However, one of the disadvantages is that they are not naturally flexible enough to capture more complex patterns.

For linear and logistic regression, the following formulae were adopted

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots \beta_{\mathrm{n}} x_{\mathrm{n}} + \mathrm{e} \quad (2)$$

$$Ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \cdots \beta_{\mathrm{n}} x_{\mathrm{n}} \quad (3)$$

Where y_i represents either a vector of values predicted (For this study can be the number of exotic animals and the amount of milk produced). $Ln\left(\frac{P}{1-P}\right)$ is a logit of y as a response i.e. for this study was a breeding method to be used by a farmer and the use of concentrate. "β" are vectors of coefficients associated with each explanatory variable, "x" are incidence matrices that link the fixed effects of the explanatory variable and *e* is error term. Decision variables were modeled as a function of explanatory variables (breeding method, purchase of concentrate, animal productivity and number of exotic animals).

(ii) Decision tree

Decision trees (DT) are among the simplest, most intuitive, easily interpretable, and widely used machine learning algorithms. They are non-parametric and therefore do not require normality assumptions of the data (Khare *et al*., 2017), can handle data of different types including continuous, categorical, ordinal, and binary hence transformations of the data are not required. It builds a classification or regression model in the form of a tree structure by evaluating the information gain of each feature (i.e., independent variable). Decision tree has been widely used to identify target groups and potential interactions of different factors (Chickering & Heckerman, 2000) i.e. looking for best potential customers, predict outcomes, data exploration, and pattern detection.

The current study defined prediction rules for a classification problem based on recursive partitioning by conditional inference (Hothorn, Hornik & Zeileis, 2006). This technique uses permutation tests and statistically determines which variables are most important and how the splits are made. It created a split by choosing the most informative feature which divides the records into left and right nodes of the tree. This technique involved three steps:

a) Test for the global null hypothesis of independence between any of the input variables and the response. It stops when the hypothesis cannot be rejected, otherwise, it continues to select the input variable with the strongest association to the response. Therefore, the m-dimensional covariate vector was defined as a vector of

$$\mathrm{X} = x_1, x_2, x_3, x_4 \ldots . x_{\mathrm{m}} \quad (4)$$

where "X" are features selected from features selections methods

Each model was fitted based on a learning sample L_n which was defined as a

$$L_n = \{(Y_i, x_{1\mathrm{i}}, \ldots\ldots x_{\mathrm{mi})}; \mathrm{i} = 1, \ldots, \mathrm{n}\} \quad (5)$$

To create a generic algorithm recursive binary partitioning for a given learning sample L_n was formulated using non-negative integer valued case weights

$$\mathrm{X} = \mathrm{w}_1 \ldots w_{\mathrm{n}} \quad (6).$$

In each node identified by case weight w, the global hypothesis of independence was formulated in terms of the m partial hypotheses $H_0^j : \mathrm{D}\left(\frac{Y}{x_{\mathrm{j}}}\right) = D(Y)$with global null hypothesis $H_0 = \bigcap_{j=1}^{m} H_0^j$.

So, in rejecting the H_0 the model stopped the recursion and when the global hypothesis was rejected. Then measured the association between Y and each of the covariates $X_{\mathrm{j}=1}$….m by test statistics or P-values indicating the deviation from the partial hypotheses H_0^j.

b) The second step was to implement a splitting criterion in the selected input variable, where the slip was performed after maximized over all subsets of covariate selected. Then

c) Recursively repeat steps 1) and 2).

This method does not require pruning of trees comparing to other DT techniques. Also, recursive partitioning by conditional inference is not vulnerable to the so-called biased variable selection problem.

(iii) Random Forest

Random Forest is an ensemble method that uses internal bootstrapping with random feature selection to train several decision trees. One of the biggest advantages of RF is less sensitive to outlies, reduces overfitting by averaging several trees (Hothorn *et al*., 2006). Also, it is an efficient way of estimating missing values while maintaining high accuracy when a large proportion of the data is missing. Random Forests are also commonly used as feature selection methods (Genuer, Poggi & Tuleau-Malot, 2010). Based on the tree strategy, it naturally ranks features by how well they improve the purity of the node. It achieves this by creating decision trees with the greatest decrease in impurity happen at the start of the trees. Thus, by pruning trees below a particular node, it creates a subset of the most important features. This study employed RF for both tasks of feature selection and model development. This concept was also extended to Boruta which also has similar stability for feature selection as RF (Degenhardt, Seifert & Szymczak, 2017)

(iv) K-nearest neighbor (KNN)

K-nearest neighbor is a non-parametric and instance-based learning algorithm. The KNN algorithm is a robust and versatile classifier that is often used as a benchmark for more complex classifiers such as artificial neural networks (ANN) and support vector machines (SVM). Despite its simplicity, KNN can outperform more powerful classifiers and is used in a variety of applications such as economic forecasting, data compression, and genetics (Bafandeh & Bolandraftar, 2013). In the classification setting, the K-nearest neighbor algorithm boils down to forming a majority vote between the K most similar instances to a given "unseen" observation. The similarity is defined according to a distance metric between two data points. This study compared different values of K from two data sets (training and testing set) and the value of K which yield constant and maximum accuracy to both datasets were used in model development.

To predict a new value, we compute the average values of most similar farmers on the basis of the selected input attributes or features. The study used a Euclidean distance to find neighbors farmers which are given by

$$d_i = \sqrt{\sum_{j=1}^{x} \Delta a_{ij}^2} \quad (7)$$

Where d_i is the distance of the ith=K farmers from the targeted farmer and Δa_{ij} is the difference of the ith=K farmers from the targeted farmer attributes and x are number of features that were used to develop a model.

(v) Artificial Neural network and Gaussian Mixture model (GMM)

Artificial Neural network is an information processing paradigm that works like a biological nervous system. Artificial Neural network can be used for pattern recognition and data classification through the learning process (Schmidhuber, 2015). It is flexible in changing environment and can handle very complex interactions. Hence, it can be used to model data which is too difficult to model with traditional approaches (Grossberg, 2017; Loyola, Pedergnana & Gimeno, 2016). While GMM is a statistical model that describes spatial distribution and characteristics of the data by assuming it can be represented as a mixture of normal (Gaussian) distributions (Scrucca, Fop, Murphy & Raftery, 2016).

3.4.3 Models evaluation

Each model was evaluated with variables sets of features selected by the different feature selection algorithm. Models were evaluated using prediction accuracy for classification problems (use or not use concentrate and breeding method: AI or bull) and R^2 values for numeric predictions (milk productivity and the number of exotic animals). We also adopted AUROC (Area Under the Receiver Operating Characteristics) as evaluation metrics for checking the classification model's performance. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. Generally, is plotting True Positive Rate (TPR) against False Positive Rate (FPR).

TPR is defined as:

$$TPR = \frac{TP}{TP + FN} \quad (8)$$

And FPR is defined as:

$$FPR = \frac{FP}{PP + TN} \quad (9)$$

Where TP = true positive, TN = true negative, FP = false positive, FN = false negative.

The study adopted K-fold cross-validation for both problem (classification and regression). Cross-validation was performed to generalizes predictive ability and avoids the problem of over-fitting which may arise while developing a model. Models accuracy was calculated through a stratified k-fold cross-validation approach (with k = 10). Data were split into k sections or 'folds' of approximately equal size that allows a model to be evaluated for each fold. Therefore, in this study overall model prediction performance of classification problem was calculated as the mean value across all folds using the following formula:

$$Averageaccuracy = \frac{\sum_i^k \frac{tp_i+tn_i}{tp_i+fn_i+fp_i+tn_i}}{k} \quad (10)$$

Where tp_i= Number of items correctly identified as positive at k_i iteration, tn_i= Number of items correctly identified as negative at k_i iteration, fn_i= Number of items wrongly identified as negative at k_i iteration and fp_i= Number of items wrongly identified as positive at k_i iteration.

Adjusted R^2 was adopted as an evaluation technique for regression problems (predict the number of exotic animals and the amount of milk to be produced). Adjusted R^2 is an extension of R^2 and was implemented based on its ability to measure model performance while accounting for the number of terms (variables) in a model. In this study the following formula was adopted:

$$R^2{}_{adj} = \frac{\sum_i^k 1 - \left[\frac{(1-R^2)(n-1)}{n-p-1}\right]}{k} \quad (11)$$

where $R^2 = 1 - \frac{\sum_{i=1}^n (Y-\widehat{Y})^2}{(Y_i-\overline{Y}_i)^2}$ (12)

where $(Y-\widehat{Y})^2$ is an average of the squares of the residuals,$(Y_i-\overline{Y}_i)^2$ account for the variance in Y values, n=number of observations, p is the number of predictors and k are the number of K-fold iteration. Therefore, the model's performance for regression problems was obtained by averaging adjusted R-squared from k-fold cross validation; for this study (k=10).

3.4.4 Models validation

The validation process was performed using Rwanda data that was collected in GIRINKA program. GIRINKA is a program that was initiated for the aim of increasing household incomes of poor farmers in Rwanda (Mutarutwa, 2014). Where each poor family was given one cow. Therefore, this study employed machine learning models to predict the sustainability of the program and the types of techniques that farmers will adopt on the farm. The study used algorithms that were selected to develop models for other countries as one way of testing the robustness of the models and if the models were overfitting. Three decisions were tested that are

(i) If a farmer will continue to keep the GIRINKA animal, which guarantee the sustainability of the program,
(ii) Whether a farmer will prefer to supplement their animals and
(iii) The breeding methods to be used by a farmer: Artificial Insemination (AI) or natural method (bull).

The validation set contains data from 1564 farmers who are beneficiaries of GIRINKA program. The model development process followed all procedures required for model development as elaborated in preceded sections.

Two decisions (response variables) to be modeled were binary; Farmer to supplement or Not and whether a farmer will use Artificial insemination or Natural bull. The other response variable was categorical measures with four options; If a farmer will continue to keep the animal or pass the animal to other family members or sell them to get income or use it for other purposes.

CHAPTER FOUR

RESULTS AND DISCUSSION

4.1 Characterizations of decision making by small scale dairy farmers

To enhance productivity and realize genetic gain, robust and practical germplasm delivery technologies and mechanisms such as artificial insemination (AI) and selective bull mating are fundamental. Since its introduction 60 years ago, AI has experienced rapid diffusion and usage across the world due to its potential. Its appeal lies not only in its ease to obtain genetic improvement but also in the elimination of costly venereal diseases, increased efficiency of bull usage which decreases running cost (Foote, 1996). It has been the most widely used reproductive technology in dairy farming and has been mainly adopted in developed countries and on commercial farms in developing countries (Chupin *et al*., 1995).

Currently, AI is also well-utilized in some African countries such as South Africa and Kenya (Bayerni, 2012). However, the adoption rate in other SSA countries is still low especially among small scale farmers (Mugisha *et al*., 2014; Tefera *et al*., 2014). The reasons for the low uptake of AI by farmers have never been clearly known across the main dairying countries in Africa. Understanding the key drivers of a farmer's choice for a particular breeding service is critical if the adoption rates are to be increased (Murage *et al*., 2011). With respect to AI, in order to formulate relevant breeding policies, there is a need of understanding the key drivers of production and farmers preference for a particular breeding service. Therefore, the study sought to determine the following: (a) whether there are any observable differences in the usage of AI and bull service among small-hold farmers in the four SSA countries, (b) the factors that influence farmer's decision on usage of AI or bull service, and (c) The similarities and differences of the main drivers for farmer's decision making with regards to the breeding method.

In summary, the results showed that there was a significant difference in animal husbandry practices between farmers who used artificial insemination (AI) and those who practiced bull mating. The majority of farmers who used AI kept records, purchased more animal feeds, had more labor by hiring workers whose average wages were higher than those of bull service farmers. However, farmers who used AI pay more for services such as water access and

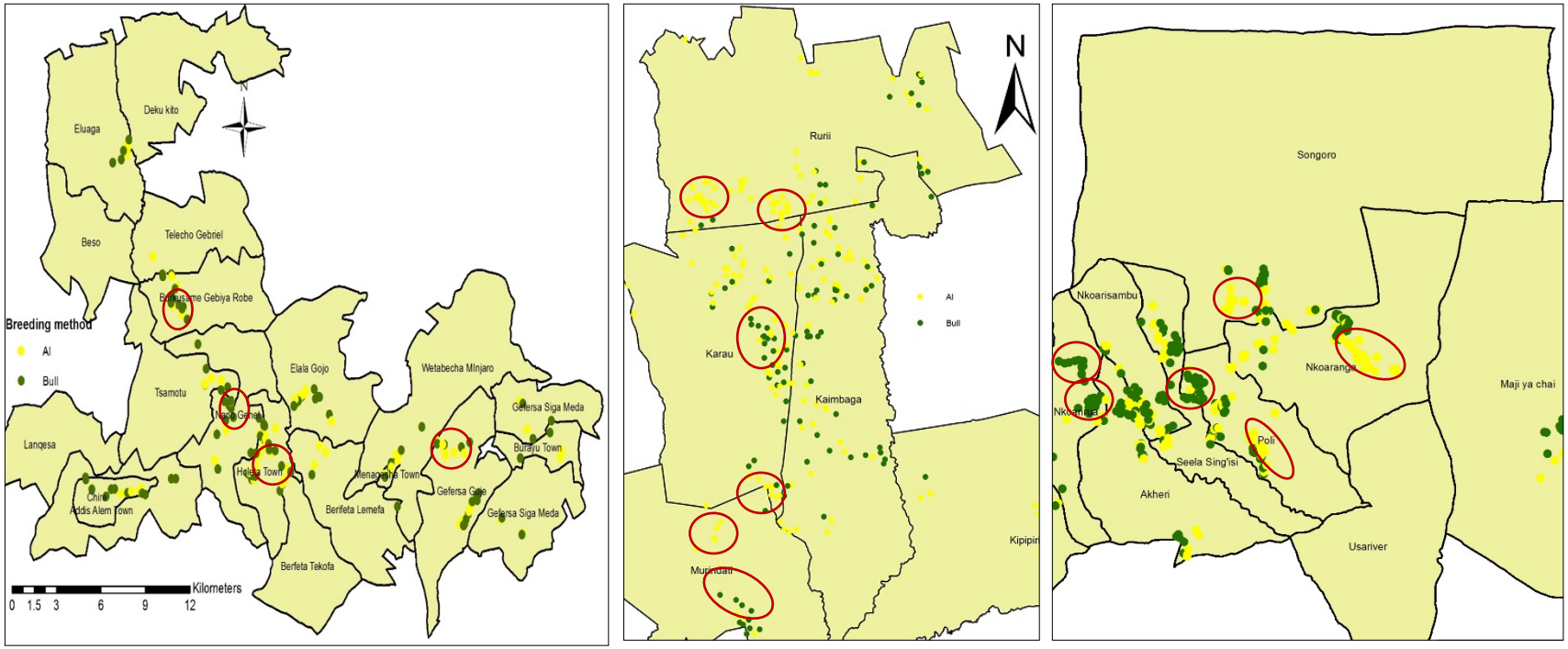
breeding while their service providers had to cover long distances compared to farmers who used bulls. This indicates limited access to services and service providers for AI farmers.

The proportion of AI to bull service users was even for Ethiopia and Kenya, while in Uganda and Tanzania, more farmers preferred bull service to AI. It was established that there were several factors that influence farmers' breeding decisions which were not the same across the regions. Factors such as farmer's experience in dairy farming, farmer's ability to keep records, and management practices such as water provision and availability of feeds had a significant association ($p < 0.001$) with AI adoption among dairy farmers. While having a large herd and large land size negatively influenced AI adoption.

It was interesting to note that farmers themselves can have influence with each other on the decisions to be made on the farm. That irrespective of the breeding method utilized, most farmers (80%) did not belong to any farmers' groups, even though these groups existed. Although, there was a clear grouping of farmers into spatial clusters as shown in Fig. 12 which was repetitive to all the countries. This indicated the influence of neighbors in dairying. These clusters coincided with the preferred method of breeding given that farmers in very close clusters tended to choose similar methods of breeding.

On institutional settings, cost of AI service and the distance covered by the service provider negatively affected ($p < 0.001$) the choice of AI as a breeding option. It was also observed that the number of services before conception was negatively associated with choice of AI as a breeding method in Ethiopia, Tanzania, and Uganda. Though factors, such as the breeding method that led to recently calved, the number of times a farmer had used AI (frequency of using AI), and accessibility to breeding method, all had a positive significant association to the breeding decision the farmer adopted in all countries. It was also concluded that factors that influence farmers' decisions were not the same across regions. Thus, one solution does not fit all.

More details on this objective is summarized in the study that was done by Mwanga *et al.* (2018) which is also attached to the list of appendixes.



(a) Ethiopia (c) Kenya (b)Tanzania

Figure 12: Clusters of farmers based on their breeding method preferences for Ethiopia, Kenya and Tanzania. The pattern is replicated in all study sites and all countries. Green dots represent farmers who used traditional bull mating while yellow dots represent farmers who used artificial insemination

4.2 Machine learning models for predicting the use of different animal breeding services in smallholder dairy farms in Eastern Africa

In the previous section, we saw how different factors can influence farmers' decisions. Hence in this section, an intensive feature selection process was performed to identify key features that could be used in model development.

Three different features selection methods namely Logistic Regression (LR), Random Forest (RF) and Boruta Algorithm (BA) where comparatively used to extract a unique set of features from more than 120 variables. All sets of variables selected from the features selection models were used in the process of developing predictive models using six algorithms including Neural Network (NN), Logistic, K-nearest Neighbor (KNN), Decision Tree (DT), Random Forest (RF) and Gaussian Mix-Model (GMM). The model performances were compared against each set of features to find the most robust models, respective for each country as demonstrated in Fig. 13. This approach was used as a pilot study to understand how models work and use the experience gained to model other decisions.

4.2.1 Features selection

Various features were selected to be important by the three features selection algorithms (RF, BA, LR) and were then used to develop the models. To gain an understanding of how features were selected, we have summarized the top ten most important variables as shown in Table 4. A complete list of features selected is summarized in appendix 3 (Table 20, Fig. 33-35)

Each feature selection algorithm had its own preference. Some features which were considered to be the most significant to one algorithm were not considered to be significant for another. Also, each feature was ranked differently except for some variables. Two features ("breeding method recently calved" and "number of times a farmer had used AI methods") were considered to be the most important predictors to all the countries.

In Ethiopia, the top ten features selected by logistic regression were variables (in order of importance) -{4, 1, 5, 8, 6, 13, 3, 9, 2, 14}, while random forest selected - {4, 5, 8, 6, 11, 1, 15, 18, 10, 14}, and Boruta selected {4,5,8,11,6,1,10,15,9,3}. Kenya had the following top ten features selected by logistic regression {4, 5, 8, 4, 3, 9, 11, 37, 22, 23}, random forest -{4, 2, 22, 11, 8, 10, 19, 18, 26, 37}, and Boruta algorithm -{4, 5, 11, 8, 22, 10, 27, 3, 26, 13}.In Tanzania, the top ten features selected by logistic regression were {4, 8, 29, 9, 11, 30, 5, 25,

23, 2), by random forest- {7, 8, 4, 5, 11, 10, 6, 9, 1, 19}, and by Boruta algorithm – {1, 36, 5, 7, 11, 4, 12, 28, 30, 8). Uganda had the following top ten features selected by logistic regression -{4, 5, 18, 21, 25, 24, 32, 17, 31, 34}, random forest – {4, 5, 33, 16, 25, 17, 10, 11, 18, 19}, and by Boruta algorithm – {4, 5, 10, 11, 16, 33, 20, 17, 18, 35}.

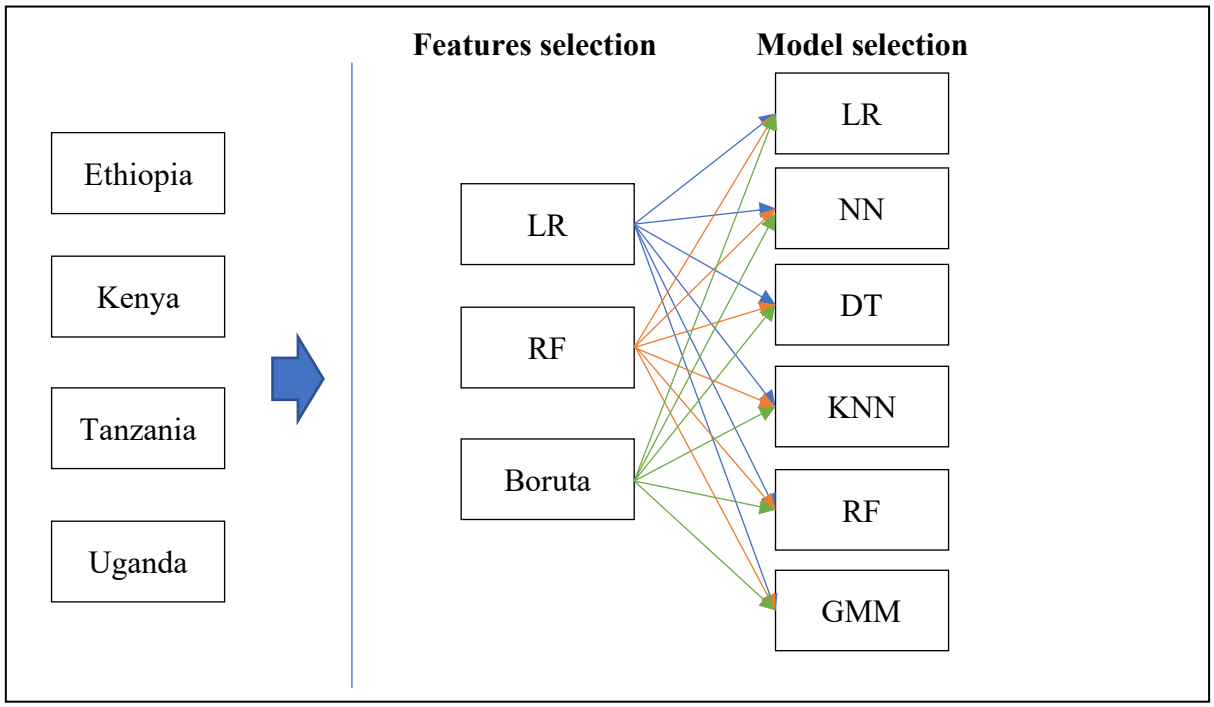


Figure 13: Approach used for features and model selection

4.2.2 Model selection

Each set of important variables selected by the feature selection method was used to develop prediction models. The feature selection techniques were evaluated using classification accuracy.

Table 5 shows the model performance (accuracy) and execution time taken for each algorithm and Fig. 14-17 indicates the ROC, also as one way to measure models' performance. Generally, the models had high predictive power (Range from 67% to 95%). However, there was a slight change in the accuracy of some models. A difference of 1 to 35% was observed for a different set of features. The neural network and GMM were very sensitive to the change of features. On execution time DT maintained low execution time than other models while NN executed for a long time than other models.

Table 4: Shows the top ten important variables selected by features selection methods: logistic (L), Random forest (R) and Boruta (B) respective for each country (Ethiopia, Tanzania, Kenya and Uganda). The numbers show variable importance

		Ethiopia			**Kenya**			**Tanzania**			**Uganda**		
S/No	**Variable**	**L**	**R**	**B**	**L**	**R**	**B**	**L**	**R**	**B**	**L**	**R**	**B**
1	Study sites	2	6	6	5				9	1			
2	Belong to farmer groups (Y/N)	9				2		10					
3	Find preferred service (Y/N)	7		10	5		8						
4	Breeding method recently calved (AI/Bull)	1	1	1	1	1	1	1	3	6	1	1	1
5	Number of times have used AI	3	2	2	2		2	7	4	3	2	2	2
6	Service provided by the government (Y/N)	5	4	5					7				
7	Service provided by private (Y/N)								1	4			
8	Service provided by individual (Y/N)	4	3	3	3	5	4	2	2	10			
9	Service provided by farmer himself (Y/N)	8		9	6			4	8				
10	Distance to service provider		9	7		6	6		6			7	3
11	Average cost for breeding		5	4	7	4	3	5	5	5		8	4
12	Amount of money spent to purchase water									7			
13	Number of months purchased fodder	6					10						
14	Grow cash crops	10	10										
15	Number of months purchased crop residue		7	8									
16	Feeding system used during dry seasons											4	5
17	Feeding system used during rain seasons										8	6	8
18	Milk production		8			8					3	9	9
19	Animal lactation length					7			10			10	
20	Frequency for watering animals												7

21	Service provided by the cooperation (Y/N)			4						4		
22	Keep breeding records (Y/N)			9	3	5						
23	Use records for animal identity (Y/N)			10			9					
24	Use records for self-evaluation (Y/N)									6		
25	Asked by extension officer to keep records (Y/N)						8			5	5	11
26	Use records for traceability (Y/N)				9	9						
27	Number of bull animals					7						
28	Water animals using tap water (Y/N)							8				
29	Water animals from the river (Y/N)						3					
30	Distance travelled to watering sources	12					6	9				
31	Frequency of treating animals									9		
32	Grow food crops									7		
33	Keep health records (Y/N)										3	6
34	Keep growth records (Y/N)									10		
35	Use records for self-evaluation (Y/N)											10
36	Availability of vaccination service (Y/N)							2				
37	Use animal tags (Y/N)			8								
38	Keep records (Y/N)				10							

Table 5: Models accuracies and time taken for each model to execute

Ethiopia

	NN		Logistic		KNN		Decision tree		Random forest		GMM	
	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)
Logistic (18)	87.39	24.09	87.72	4.24	88.15	6.13	85.85	0.247	89.58	12.19	86.95	7.72
Random forest (16)	53.13	117.42	89.4	4.79	88.92	5.12	89.6	0.27	90.49	11.52	87.647	7.86
Boruta (12)	87.3	69.9	88.65	4.37	90.8	5.39	88.87	0.22	90.56	10.55	86.62	1.27

Kenya

	NN		Logistic		KNN		Decision tree		Random forest		GMM	
	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)
Logistic (23)	52.98	90.66	93.88	9.39	92.9	16.27	93.14	0.56	94.63	23.37	89.26	5.72
Random forest (9)	43.53	25.53	94.02	6.73	93.4	2.9	92.82	0.247	93.57	42.17	92.97	6.08
Boruta (17)	65.57	94.2	94.15	7.72	93.4	10.15	92.85	0.27	94.35	18.67	91.97	4.426

Tanzania

	NN		Logistic		KNN		Decision tree		Random forest		GMM	
	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)
Logistic (17)	54.11	40.38	95.28	56.467	95.08	10.38	94.67	0.47	94.67	14.97	89.65	14.15
Random forest (13)	94.02	12.26	93.07	4.418	92.9	5.69	93.93	0.27	94.37	9.84	92.29	1.84
Boruta (17)	73.42	35.052	92.31	5.55	92.5	6.65	92.22	0.28	93.61	11.47	91.48	51.27

Uganda

	NN		Logistic		KNN		Decision tree		Random forest		GMM	
	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)	Accuracy (%)	Time (seconds)
Logistic (15)	88.28	21.8	96.26	9.69	96	7.16	96.39	0.31	96.39	0.13	91.63	0.4
Random forest (16)	87.8	17.04	95.89	6.036	95	4.73	96.56	0.17	96.56	7.75	91.9	0.31
Boruta (15)	96.99	24.29	97.49	3.82	97.3	4.62	97.26	0.17	97.87	7.16	92.24	1.24

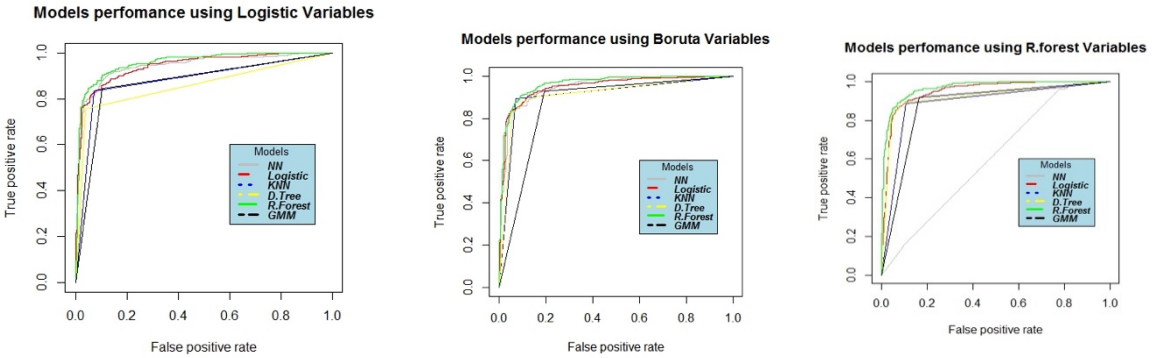


Figure 14: Models performance for Ethiopia data

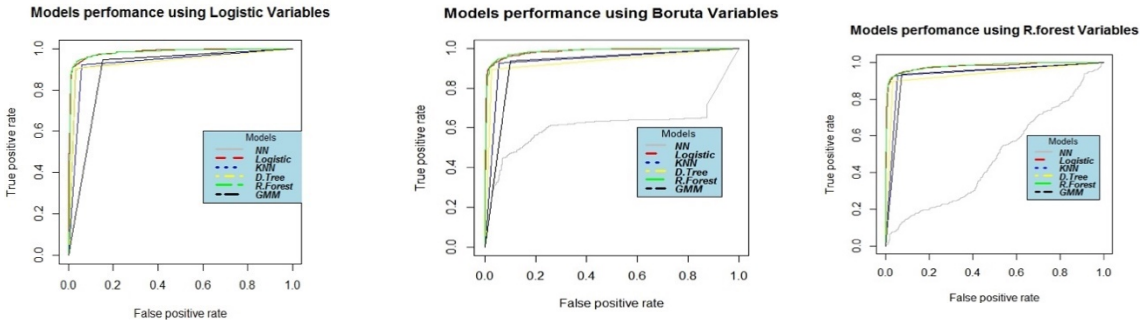


Figure 15: Models performance for Kenya data

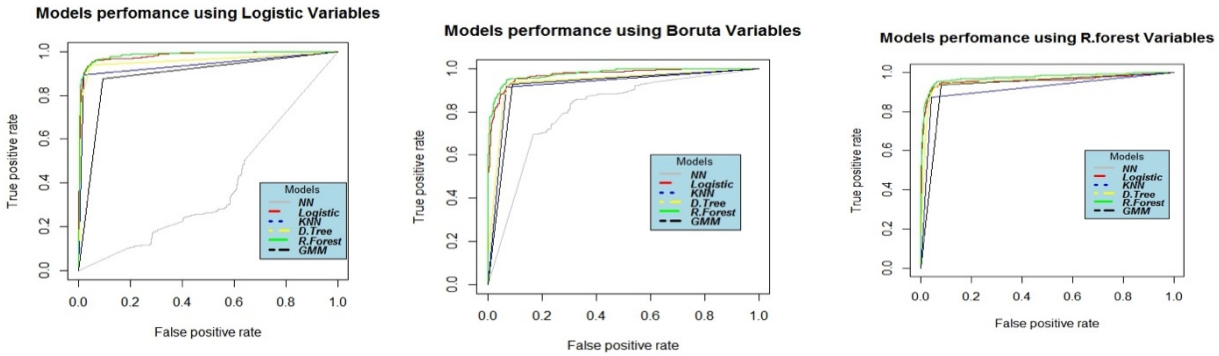


Figure 16: Models performance for Tanzania data

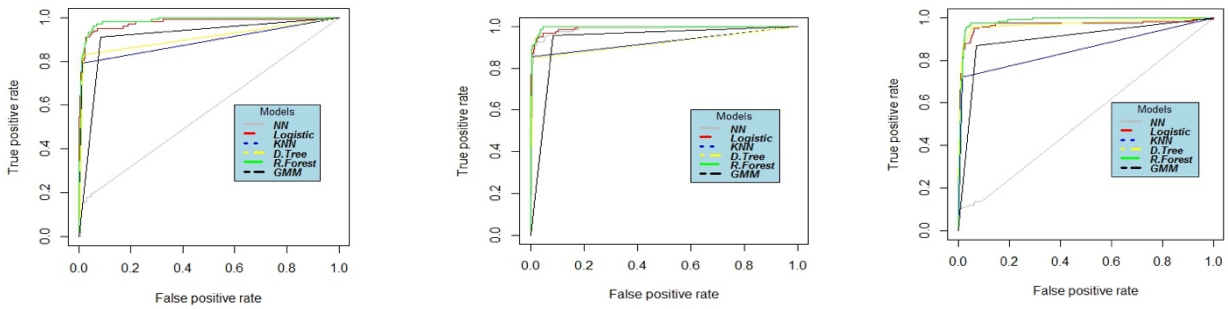


Figure 17: Models performance for Uganda data

4.2.3 Development of final, country-specific models

In selecting the algorithms that can be used in features selection and model development, we considered the following factors

(i) The model to be robust in prediction i.e. fits the data and attained high prediction accuracy
(ii) Algorithms that work for all the countries.

All models attained high accuracy except for GMM and NN which failed in some cases as described in a preceded paragraph. Therefore, the two algorithms were excluded, and remained with four options (LR, KNN, DT and RF) which all attained high accuracies. We adopted a combination of RF and DT as a feature selection method and model development. Random forest was selected for feature selection because was robust and had high performance. Random forest is considered to be a robust model based on its ability to do intensive search of features that can maximize prediction accuracy. Comparing it to Boruta which also operates on the same concept as RF, Boruta had a higher execution time than RF. The study compared the execution time for two variables using the same criteria (number of trees); i.e. at 100 iteration execution time for random forest was =44.52 seconds and BR=338.4 seconds.

Also, on model selection, DT was adopted given the fact that this study aimed at modeling decision making process for the farmer and DT naturally has been constructed specifically for modeling decisions. Because of this we found DT to be more coherent in presenting the relevant information and so relevant to our study. Its representation style gives a decision-maker alternative solutions and possible choices which make it easier to make a well-informed choice. On the other hand, DT makes good use of the 'what if' thought for decision maker to scrutinizing the possible risks and benefits that are brought about by certain choices. Additional both DT and RF accommodate nonlinear relationships compared to LR.

It was also observed that there were two variables (breeding method recently calved" and "number of times a farmer had used AI methods) which completely dominated the predictive capacity of the models. To a greater extent, the two variables dictates an obvious fact and possibility of using AI which is why they cause an increase in accuracy. Even though these variables increase the accuracy of our models they render them less useful since they are re-affirming what is already known. Therefore, in developing the final model the two features were excluded. However, dropping of these features significantly affected the model

performance for Ethiopia and Kenya. Where the accuracy for Ethiopia dropped from 90% to 81% and Kenya from 92% to 78%.

Figure 18 shows all important features that were selected by random forest after excluding the two variables. Though Not all variables that were selected by random forest were used in the final model development. Decision tree algorithm also performed a univariate analysis for the entire set of variables and selects the input variable that best separates the data with respect to the class variable (Use of AI of Bull). After screening, the DT used the most significant variables which were verified by the gain ratio. Nonetheless, the variables selected by decision trees were also ranked higher by random forest.

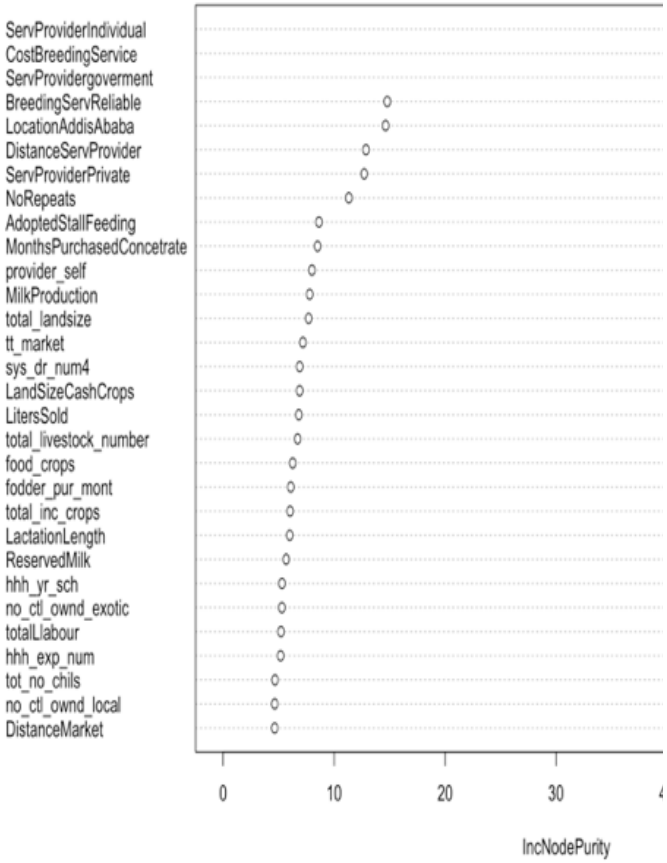
4.2.4 Models to predict the adoption of AI as a breeding method

The final country-specific models were able to classify the previously unseen sets of data (testing sets) at the accuracy of 81% for Ethiopia, 78% for Kenya, 93% for Tanzania, and 90% for Uganda.

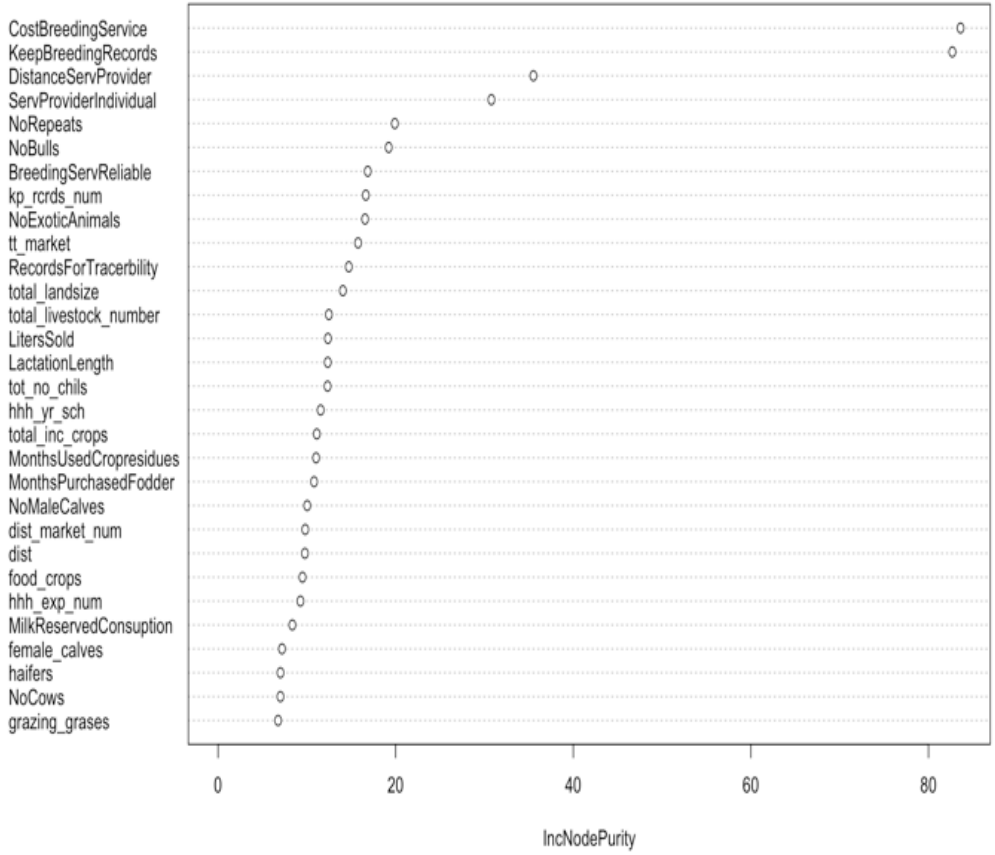
In Ethiopia out of 15 variables, 11 were selected by a DT algorithm to build a model with 37 (Inner and terminal) nodes which attain prediction accuracy of 81%. The final form of the decision tree model is shown in Fig. 19 and Table 6.

The top three key drivers that influence farmers' decisions in regard to the breeding method to be adopted were, type of service provider (where farmers can access the service), farm location (urban or rural) and feeding systems that have been adopted on a farm. The results showed that a specific combination of factors in the farmer profile determined the breeding service they could go for. For example, farmers who had a chance of accessing breeding services from the government, are located in Addis Ababa and preferred stall feeding as their feeding system can access breeding service even though have to repeat more than once for a successful conception. These farmers were more likely (at 92% accuracy) to adopt AI (Node 37).

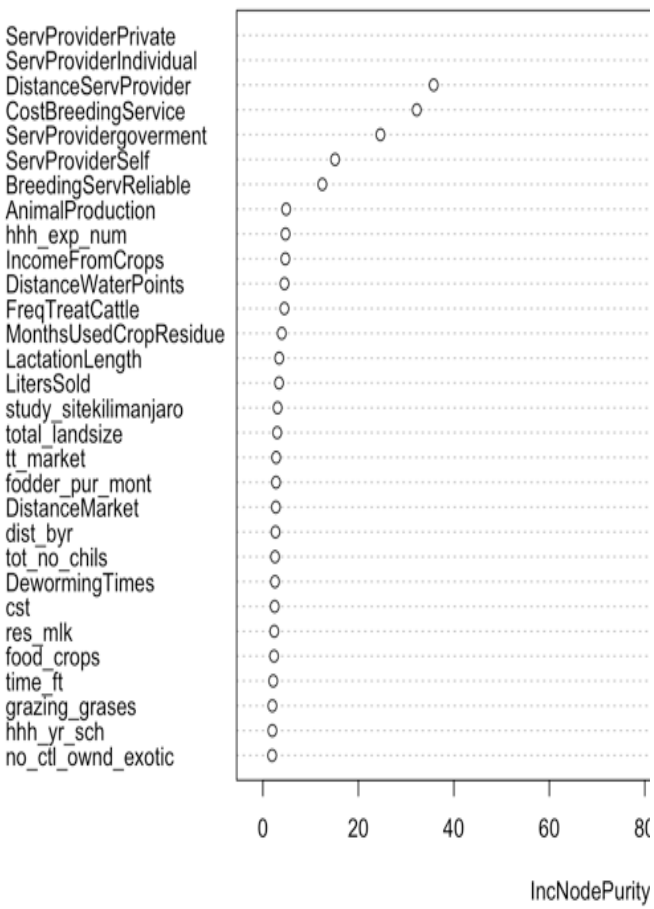
The results also showed that the strength of the farmers' agronomic orientation played a major role. For example, farmers with a small landholding area (<1.75 acres), which was intensively used for cash crops and their animals managed to attain high production (10 liters/animal/day) were more likely to use AI (Node 23).



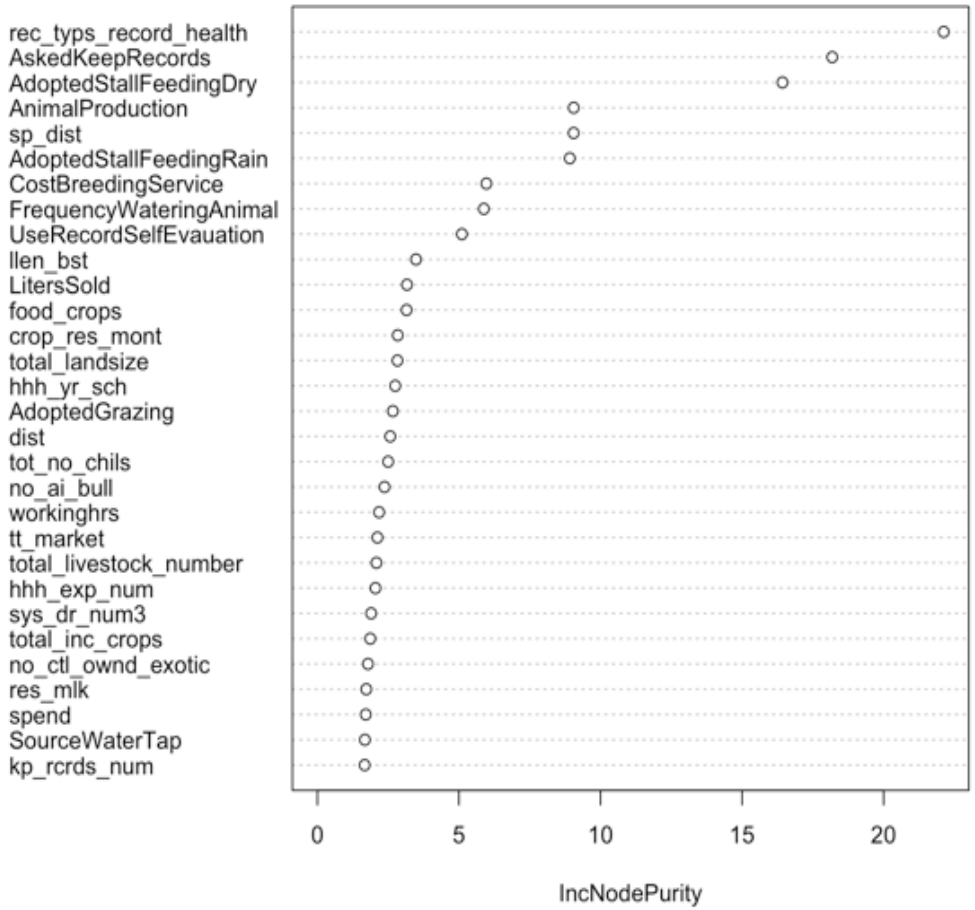
a) Ethiopia



b) Kenya



c) Tanzania



d) Uganda

Figure 18: Variables selected by Random Forest for each country respectively.

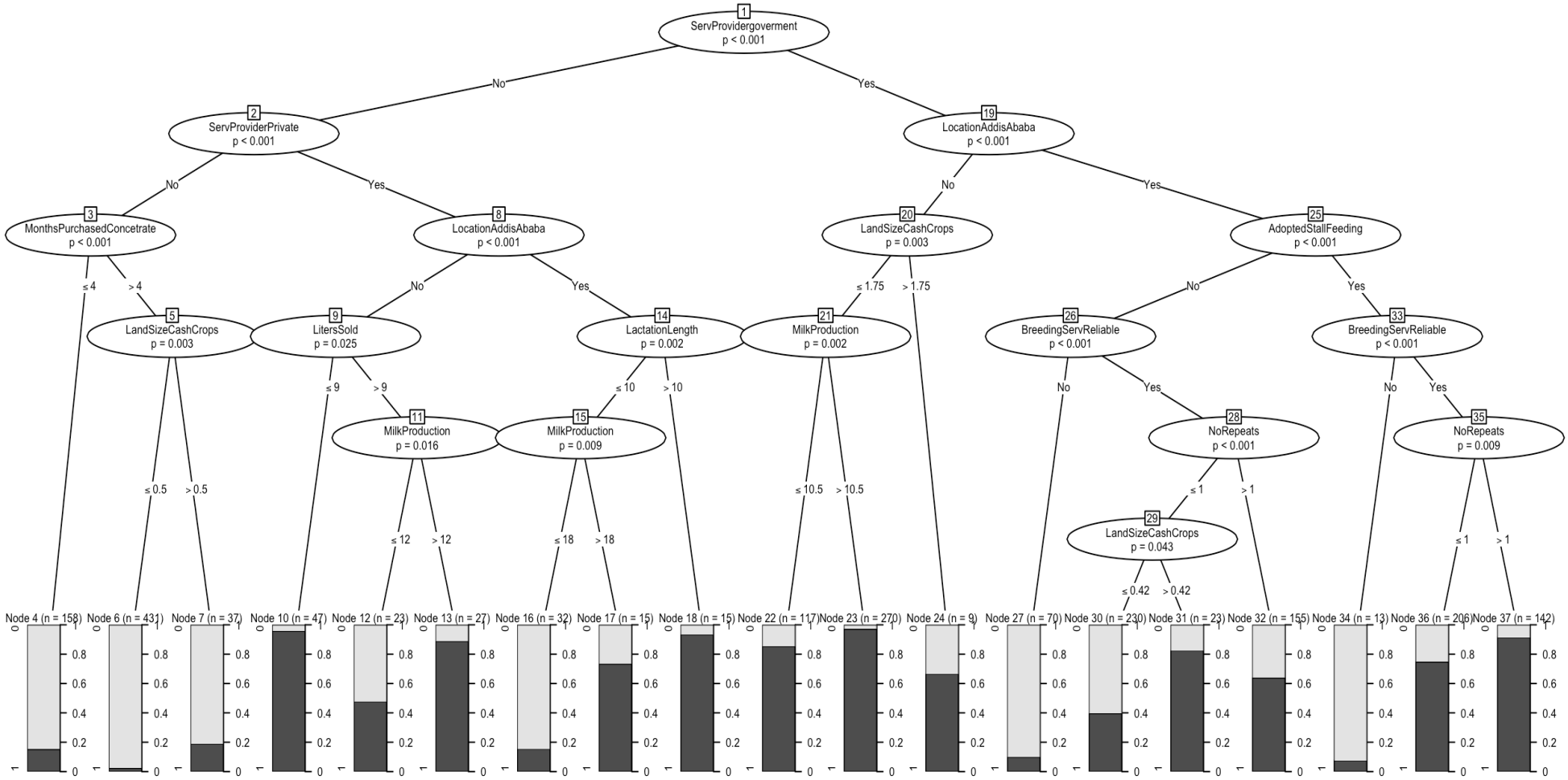


Figure 19: Decision tree model for Ethiopia

Table 6: Summary of decision tree model for predicting farmers decisions in regard to the AI adoption: Ethiopia

Fitted DT model for Ethiopia

```
[1] root
|  [2] ServProvidergoverment in No
|  |  [3] ServProviderPrivate <= 0
|  |  |  [4] MonthsPurchasedConcetrate <= 4: 0 (n = 158, err = 15.1899%)
|  |  |  [5] MonthsPurchasedConcetrate > 4
|  |  |  |  [6] LandSizeCashCrops <= 0.5: 0 (n = 431, err = 2.3202%)
|  |  |  |  [7] LandSizeCashCrops > 0.5: 0 (n = 37, err = 18.9189%)
|  |  [8] ServProviderPrivate > 0
|  |  |  [9] LocationAddisAbaba in No
|  |  |  |  [10] LitersSold <= 9: 1 (n = 47, err = 4.2553%)
|  |  |  |  [11] LitersSold > 9
|  |  |  |  |  [12] MilkProduction <= 12: 0 (n = 23, err = 47.8261%)
|  |  |  |  |  [13] MilkProduction > 12: 1 (n = 27, err = 11.1111%)
|  |  |  [14] LocationAddisAbaba in Yes
|  |  |  |  [15] LactationLength <= 10
|  |  |  |  |  [16] MilkProduction <= 18: 0 (n = 32, err = 15.6250%)
|  |  |  |  |  [17] MilkProduction > 18: 1 (n = 15, err = 26.6667%)
|  |  |  |  [18] LactationLength > 10: 1 (n = 15, err = 6.6667%)
|  [19] ServProvidergoverment in Yes
|  |  [20] LocationAddisAbaba in No
|  |  |  [21] LandSizeCashCrops <= 1.75
|  |  |  |  [22] MilkProduction <= 10.5: 1 (n = 117, err = 14.5299%)
|  |  |  |  [23] MilkProduction > 10.5: 1 (n = 270, err = 2.9630%)
|  |  |  [24] LandSizeCashCrops > 1.75: 1 (n = 9, err = 33.3333%)
|  |  [25] LocationAddisAbaba in Yes
|  |  |  [26] AdoptedStallFeeding in No
|  |  |  |  [27] BreedingServReliable in No: 0 (n = 70, err = 10.0000%)
|  |  |  |  [28] BreedingServReliable in Yes
|  |  |  |  |  [29] NoRepeats <= 1
|  |  |  |  |  |  [30] LandSizeCashCrops <= 0.42: 0 (n = 230, err = 39.5652%)
|  |  |  |  |  |  [31] LandSizeCashCrops > 0.42: 1 (n = 23, err = 17.3913%)
|  |  |  |  |  [32] NoRepeats > 1: 1 (n = 155, err = 36.1290%)
|  |  |  [33] AdoptedStallFeeding in Yes
|  |  |  |  [34] BreedingServReliable in No: 0 (n = 13, err = 7.6923%)
|  |  |  |  [35] BreedingServReliable in Yes
|  |  |  |  |  [36] NoRepeats <= 1: 1 (n = 206, err = 25.2427%)
|  |  |  |  |  [37] NoRepeats > 1: 1 (n = 142, err = 8.4507%)
Number of inner nodes:    18
Number of terminal nodes: 19
```

While farmers who neither access breeding service from the government nor private sources and preferred to purchase concentrate for their animals (>4 months). They also had low land to grow cash crops were less likely to adopt AI as their breeding method (Node 6).

In Kenya, the main factors that characterize farmers decisions in regard to the breeding method to be adopted include farmers characteristics on husbandry practices (Whether farmers keep records), the cost for breeding service, type of service provider for breeding service and the number of bulls kept by a farmer (Fig. 20 and Table 7). Majority of farmers who were more likely to adopt AI preferred to keep records on breeding. Moreover, it was noted that farms with more number of bulls had a low probability of adopting AI. Likewise, farmers who adopted bulls were categorized to pay less for breeding service and mostly serve their animals individually. For example, farmers who were categorized in node 9, did not keep records, pay less for breeding service (<700 Kshs) and served their animals individually. They were less likely (100% accuracy) to adopt AI.

In Tanzania, the most significant factors that described farmers on their breeding choices were the type of service providers who provides breeding service to farmers, reliability of AI service and farm location (Fig. 21 and Table 8). The study found that the majority of farmers who were using AI obtained that service from both private sources and the government and the service was reliable.

Taking a case study of farmers who were classified in node 5, they neither get breeding service from private sources nor government, also were not from Tanga and often did not treat their animals against diseases. They were classified as a group of farmers with low probability (Accuracy=98.6%) of adopting AI. While those in node 24 apart from accessing the service from private sources, the service was reliable, walk short distance to access water and where not from Njombe. Their chances of adopting AI was high (92.3%).

Farm characteristics play a significant role in classifying farmers in Uganda (Fig. 22 and Table 9). Factors such as keeping of animal records (Health records) and feeding system adopted on the farm were considered as key predictors. Those who adopted AI were featured as farmers who like to keep animal records and preferred stall feeding.

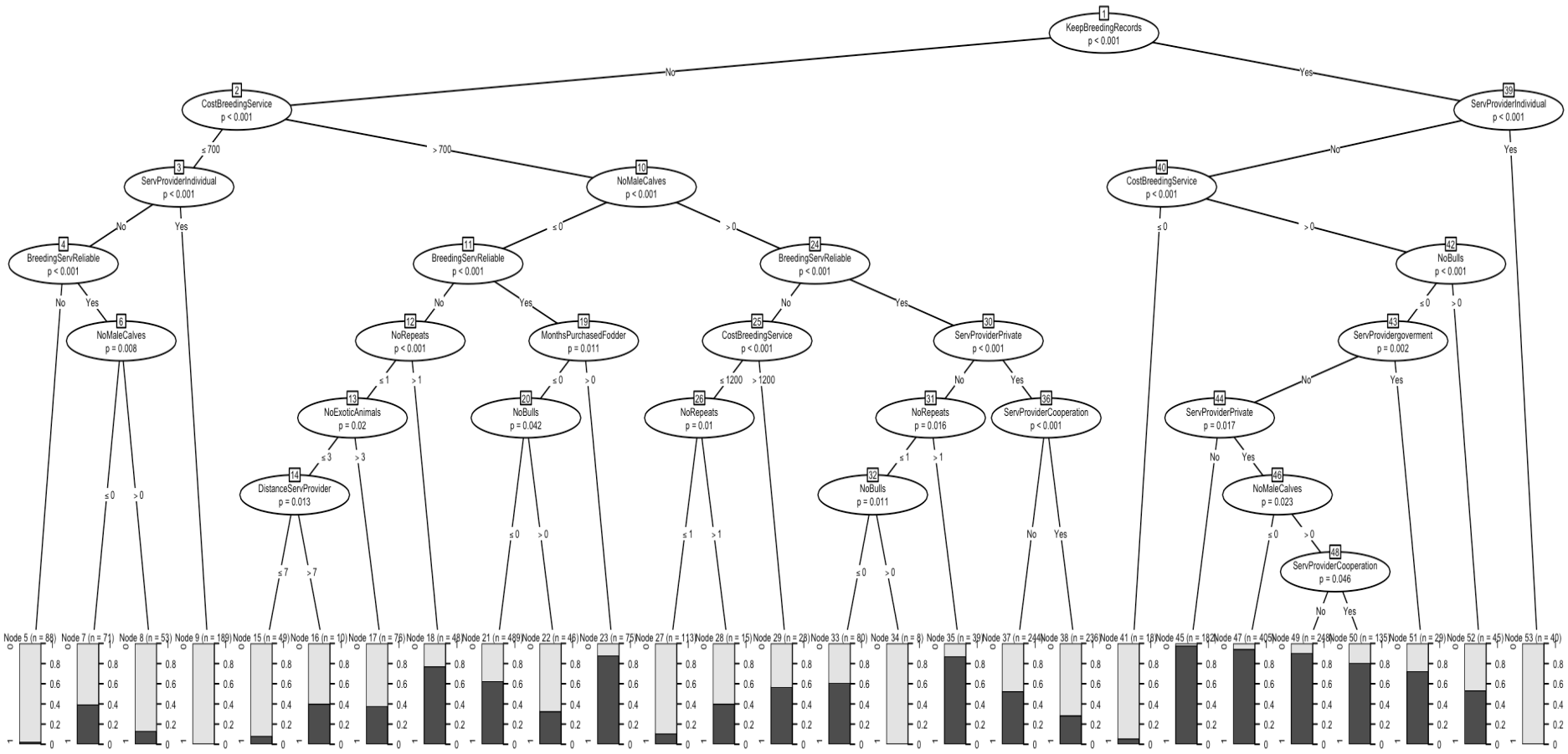


Figure 20: Decision tree model for Kenya

Table 7: Summary of decision tree model for predicting farmers decisions in regard to AI adoption: Kenya

Fitted DT model for Kenya

```
[1] root
|  [2] KeepBreedingRecords in No
|  |  [3] CostBreedingService <= 700
|  |  |  [4] ServProviderIndividual in No
|  |  |  |  [5] BreedingServReliable in No: 0 (n = 88, err = 2.2727%)
|  |  |  |  [6] BreedingServReliable in Yes
|  |  |  |  |  [7] NoMaleCalves <= 0: 0 (n = 71, err = 39.4366%)
|  |  |  |  |  [8] NoMaleCalves > 0: 0 (n = 53, err = 13.2075%)
|  |  |  [9] ServProviderIndividual in Yes: 0 (n = 189, err = 0.0000%)
|  |  [10] CostBreedingService > 700
|  |  |  [11] NoMaleCalves <= 0
|  |  |  |  [12] BreedingServReliable in No
|  |  |  |  |  [13] NoRepeats <= 1
|  |  |  |  |  |  [14] NoExoticAnimals <= 3
|  |  |  |  |  |  |  [15] DistanceServProvider <= 7: 0 (n = 49, err = 8.1633%)
|  |  |  |  |  |  |  [16] DistanceServProvider > 7: 0 (n = 10, err = 40.0000%)
|  |  |  |  |  |  [17] NoExoticAnimals > 3: 0 (n = 76, err = 38.1579%)
|  |  |  |  |  [18] NoRepeats > 1: 1 (n = 48, err = 22.9167%)
|  |  |  |  [19] BreedingServReliable in Yes
|  |  |  |  |  [20] MonthsPurchasedFodder <= 0
|  |  |  |  |  |  [21] NoBulls <= 0: 1 (n = 489, err = 37.4233%)
|  |  |  |  |  |  [22] NoBulls > 0: 0 (n = 46, err = 32.6087%)
|  |  |  |  |  [23] MonthsPurchasedFodder > 0: 1 (n = 75, err = 12.0000%)
|  |  |  [24] NoMaleCalves > 0
|  |  |  |  [25] BreedingServReliable in No
|  |  |  |  |  [26] CostBreedingService <= 1200
|  |  |  |  |  |  [27] NoRepeats <= 1: 0 (n = 113, err = 10.6195%)
|  |  |  |  |  |  [28] NoRepeats > 1: 0 (n = 15, err = 40.0000%)
|  |  |  |  |  [29] CostBreedingService > 1200: 1 (n = 28, err = 42.8571%)
|  |  |  |  [30] BreedingServReliable in Yes
|  |  |  |  |  [31] ServProviderPrivate in No
|  |  |  |  |  |  [32] NoRepeats <= 1
|  |  |  |  |  |  |  [33] NoBulls <= 0: 1 (n = 80, err = 38.7500%)
|  |  |  |  |  |  |  [34] NoBulls > 0: 0 (n = 8, err = 0.0000%)
|  |  |  |  |  |  [35] NoRepeats > 1: 1 (n = 39, err = 12.8205%)
|  |  |  |  |  [36] ServProviderPrivate in Yes
|  |  |  |  |  |  [37] ServProviderCooperation in No: 1 (n = 244, err = 47.5410%)
|  |  |  |  |  |  [38] ServProviderCooperation in Yes: 0 (n = 236, err = 28.3898%)
|  [39] KeepBreedingRecords in Yes
```

```
|  |  [40] ServProviderIndividual in No
|  |  |  [41] CostBreedingService <= 0: 0 (n = 18, err = 5.5556%)
|  |  |  [42] CostBreedingService > 0
|  |  |  |  [43] NoBulls <= 0
|  |  |  |  |  [44] ServProvidergoverment in No
|  |  |  |  |  |  [45] ServProviderPrivate in No: 1 (n = 182, err = 2.1978%)
|  |  |  |  |  |  [46] ServProviderPrivate in Yes
|  |  |  |  |  |  |  [47] NoMaleCalves <= 0: 1 (n = 405, err = 5.6790%)
|  |  |  |  |  |  |  [48] NoMaleCalves > 0
|  |  |  |  |  |  |  |  [49] ServProviderCooperation in No: 1 (n = 248, err = 9.2742%)
|  |  |  |  |  |  |  |  [50] ServProviderCooperation in Yes: 1 (n = 135, err = 19.2593%)
|  |  |  |  |  [51] ServProvidergoverment in Yes: 1 (n = 29, err = 27.5862%)
|  |  |  |  [52] NoBulls > 0: 1 (n = 45, err = 46.6667%)
|  |  [53] ServProviderIndividual in Yes: 0 (n = 40, err = 0.0000%)

Number of inner nodes:  26
Number of terminal nodes: 27
```

For example, farmers in node 23, kept animal records, adopted stall feeding although they had to inseminate their animals more than once for a successful conception rate. They were classified with high probability (Accuracy 96.3%) of adopting AI. While those in node 7, did not keep records, use other feeding systems apart from stall feeding, access water from other sources, they give less water to their animals and often did not use crop residue. Their chances of adopting AI was very low.

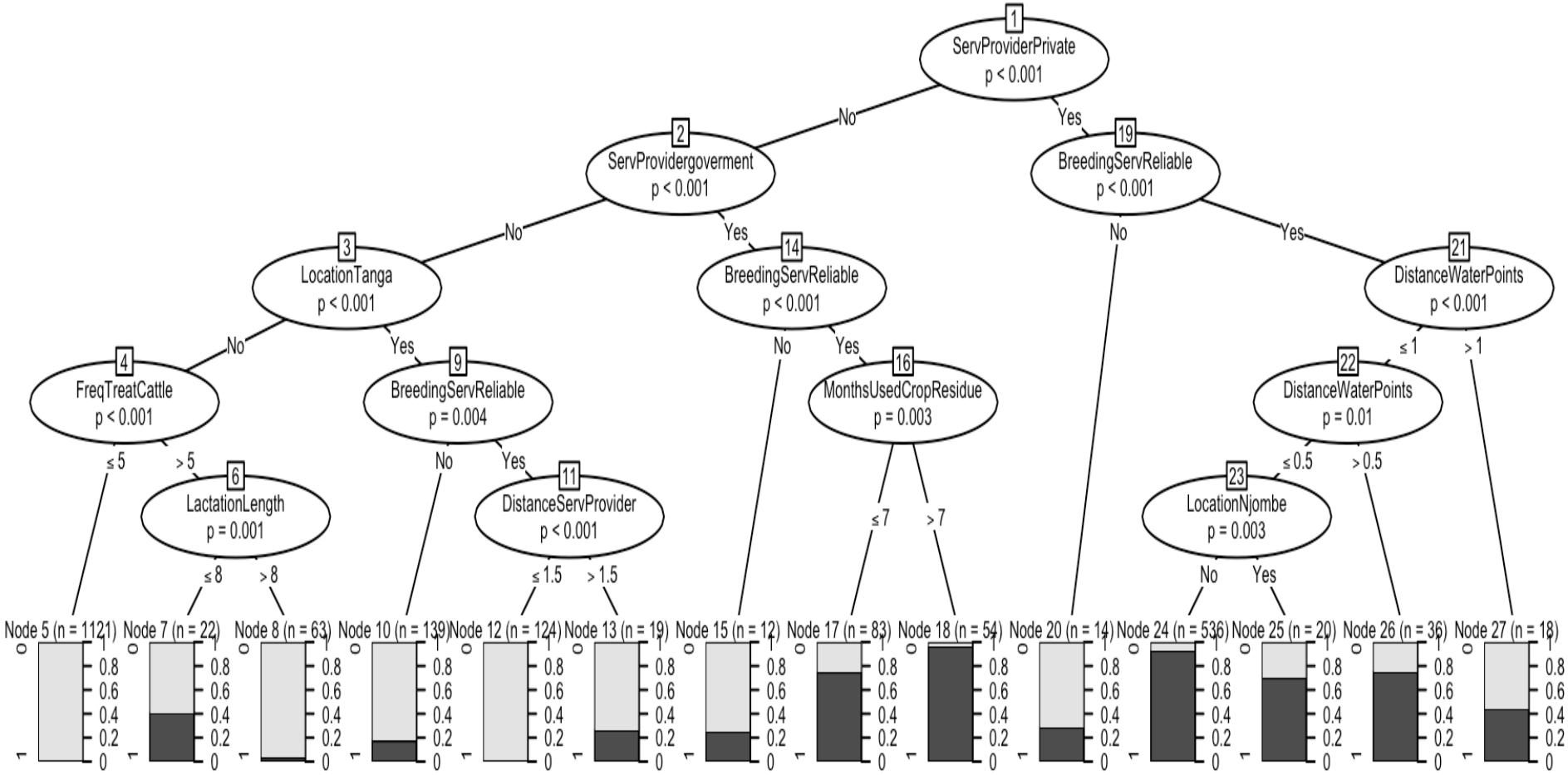


Figure 21: Decision tree model for Tanzania

Table 8: Summary of decision tree model for predicting farmers decisions in regard to AI adoption: Tanzania

```
Fitted DT model for Tanzania:
[1] root
|  [2] ServProviderPrivate in No
|  |  [3] ServProvidergoverment in No
|  |  |  [4] LocationTanga in No
|  |  |  |  [5] FreqTreatCattle <= 5: 0 (n = 1121, err = 1.4273%)
|  |  |  |  [6] FreqTreatCattle > 5
|  |  |  |  |  [7] LactationLength <= 8: 0 (n = 22, err = 40.9091%)
|  |  |  |  |  [8] LactationLength > 8: 0 (n = 63, err = 3.1746%)
|  |  |  [9] LocationTanga in Yes
|  |  |  |  [10] BreedingServReliable in No: 0 (n = 139, err = 17.2662%)
|  |  |  |  [11] BreedingServReliable in Yes
|  |  |  |  |  [12] DistanceServProvider <= 1.5: 0 (n = 124, err = 0.8065%)
|  |  |  |  |  [13] DistanceServProvider > 1.5: 0 (n = 19, err = 26.3158%)
|  |  [14] ServProvidergoverment in Yes
|  |  |  [15] BreedingServReliable in No: 0 (n = 12, err = 25.0000%)
|  |  |  [16] BreedingServReliable in Yes
|  |  |  |  [17] MonthsUsedCropResidue <= 7: 1 (n = 83, err = 25.3012%)
|  |  |  |  [18] MonthsUsedCropResidue > 7: 1 (n = 54, err = 3.7037%)
|  [19] ServProviderPrivate in Yes
|  |  [20] BreedingServReliable in No: 0 (n = 14, err = 28.5714%)
|  |  [21] BreedingServReliable in Yes
|  |  |  [22] DistanceWaterPoints <= 1
|  |  |  |  [23] DistanceWaterPoints <= 0.5
|  |  |  |  |  [24] LocationNjombe in No: 1 (n = 536, err = 7.2761%)
|  |  |  |  |  [25] LocationNjombe in Yes: 1 (n = 20, err = 30.0000%)
|  |  |  |  [26] DistanceWaterPoints > 0.5: 1 (n = 36, err = 25.0000%)
|  |  |  [27] DistanceWaterPoints > 1: 0 (n = 18, err = 44.4444%)

Number of inner nodes:    13
Number of terminal nodes: 14
```

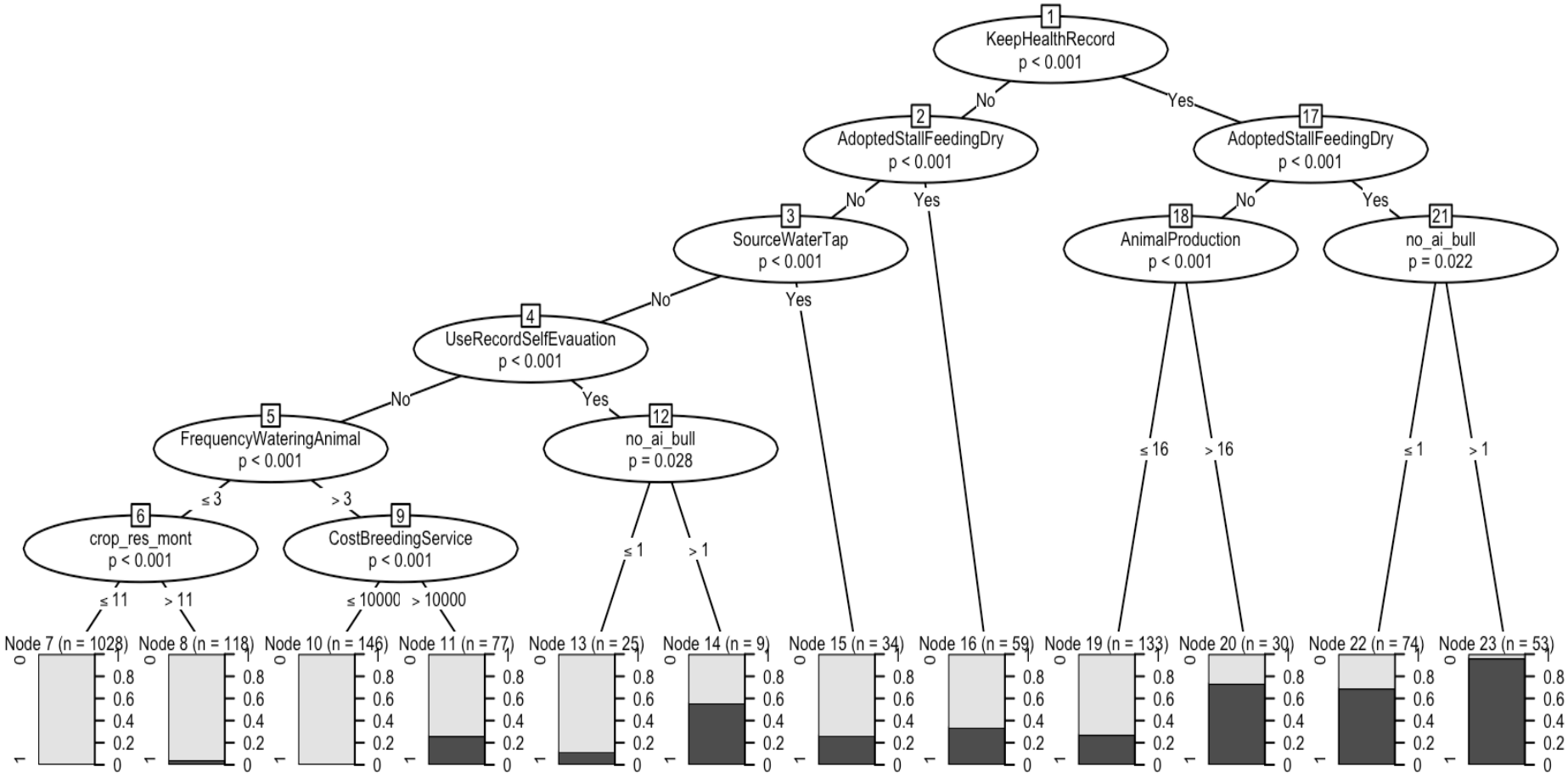


Figure 22: Decision tree model for Uganda

Table 9: Summary of decision tree model for predicting farmers decisions in regard to AI adoption: Uganda

Fitted DT for Uganda:

[1] root
| [2] KeepHealthRecord in No
| | [3] AdoptedStallFeedingDry in No
| | | [4] SourceWaterTap in No
| | | | [5] UseRecordSelfEvauation in No
| | | | | [6] FrequencyWateringAnimal <= 3
| | | | | | [7] crop_res_mont <= 11: 0 (n = 1028, err = 0.5837%)
| | | | | | [8] crop_res_mont > 11: 0 (n = 118, err = 4.2373%)
| | | | | [9] FrequencyWateringAnimal > 3
| | | | | | [10] CostBreedingService <= 10000: 0 (n = 146, err = 0.0000%)
| | | | | | [11] CostBreedingService > 10000: 0 (n = 77, err = 25.9740%)
| | | | [12] UseRecordSelfEvauation in Yes
| | | | | [13] no_ai_bull <= 1: 0 (n = 25, err = 12.0000%)
| | | | | [14] no_ai_bull > 1: 1 (n = 9, err = 44.4444%)
| | | [15] SourceWaterTap in Yes: 0 (n = 34, err = 26.4706%)
| | [16] AdoptedStallFeedingDry in Yes: 0 (n = 59, err = 33.8983%)
| [17] KeepHealthRecord in Yes
| | [18] AdoptedStallFeedingDry in No
| | | [19] AnimalProduction <= 16: 0 (n = 133, err = 27.0677%)
| | | [20] AnimalProduction > 16: 1 (n = 30, err = 26.6667%)
| | [21] AdoptedStallFeedingDry in Yes
| | | [22] no_ai_bull <= 1: 1 (n = 74, err = 31.0811%)
| | | [23] no_ai_bull > 1: 1 (n = 53, err = 3.7736%)

Number of inner nodes: 11
Number of terminal nodes: 12

4.3 Models to predict concentrate usage, keeping of exotic animals and animals' productivity

Two decisions were modeled: usage of animal supplements and keeping of exotic animals. Also, models for predicting farmers' productivity against various farm factors were developed. Based on the results obtained from the first experiment (modeling breeding decision) it was observed that ANN and GMM were not robust to our data, these algorithms were dropped at this stage. Therefore, in model selection, only four algorithms were compared these were LM for continuous variables or LR for categorical variables, DT, KNN and RF.

Similarly, the study had to identify the key predictors from a pool of more than 120 variables using features selection methods. The feature selection methods used were, LM, LR, RF and Boruta. However, in this stage, only the top 15 variables that were selected by feature selection methods were used in developing the models.

4.3.1 Models performance

Table 10,11 and Table 12 summarize results obtained on models' performance. As stated above models were built and tested based on the top 15 significant variables selected by features selection methods. This is due to the reason that each feature selection methods identified a large set of features where up to 65 features per set were selected. Therefore, to standardize the model selection process only the top 15 features were considered.

In predicting usage of concentrate, all models had a high performance with an accuracy of 90%-97% except for Kenya which was slightly lower (69%-78%) as shown in Table 10. It was generalized that in Kenya there were some of the important predictors that were missing to explain the outstanding variance.

However, the performance of these models was not the same with continuous variables including predicting the number of exotic animals and animals' productivity (Table 11 and Table 12). The major reasons for the poor performance of these models i.e. Decision Tree and Random Forest were the facts that these models work better with classification rather than Regression. Likewise, for Boruta which is built-in random forest settings.

Table 10: Models performance for predicting usage of concentrate on the farm. The results are accuracies obtained by models developed (LG: Logistic model, KNN: K-nearest neighbor, DT: Decision tree and RF: Random forest) using different set of features selected

	Ethiopia			**Kenya**			**Tanzania**			**Uganda**		
Features selections methods	LG	RF	BR	LG	RF	BR	LG	RF	BR	LG	RF	BR
LG (%)	91	90	90	78	71	71	93	93	93	93	93	93
KNN (%)	92	92	93	73	73	73	94	94	94	97	95	95
DT (%)	91	91	91	69	69	69	93	93	93	91	91	91
RF (%)	91	91	91	72	72	72	94	93	93	94	94	94

Table 11: Model performance for predicting the number of exotic animals to be kept by a farmer on the farm. The results are adjusted R^2 values obtained by models developed (LM: Linear model, KNN: K-nearest neighbor, DT: Decision tree and RF: Random forest) using different set of features selected

	Ethiopia			**Kenya**			**Tanzania**			**Uganda**		
Features selections methods	LM	RF	BR	LM	RF	BR	LM	RF	BR	LM	RF	BR
LM	0.33	0.28	0.29	0.11	0.11	0.11	0.24	0.26	0.27	0.44	0.44	0.44
KNN	0.83	0.44	46	0.95	0.78	0.69	0.66	0.55	0.66	0.77	0.57	0.57
DT	0.34	0.35	0.36	0.12	0.12	0.12	0.27	0.13	0.13	0.47	0.48	0.48
RF	0.31	0.325	0.31	0.05	0.07	0.07	0.25	0.24	0.23	0.49	0.52	0.14

Table 12: Models performance for predicting the amount of milk to be produced by the best animal. The results are adjusted R^2 values obtained by models developed (LM: Linear model, KNN: K-nearest neighbor, DT: Decision tree and RF: Random forest) using different set of features selected

	Ethiopia			**Kenya**			**Tanzania**			**Uganda**		
Features selections methods	LM	RF	BR	LM	RF	BR	LM	RF	BR	LM	RF	BR
LM	0.49	0.46	0.49	0.18	0.17	0.17	0.34	0.29	0.31	0.36	0.38	0.36
KNN	0.77	0.92	0.73	0.91	0.93	0.94	0.99	0.90	0.8	0.94	0.90	0.81
DT	0.56	0.58	0.58	0.91	0.26	0.26	0.39	0.39	0.38	0.38	0.39	0.38
RF	0.54	0.58	0.58	0.22	0.26	0.26	0.38	0.4	0.39	0.37	0.39	0.37

For classification problems, all models attained high accuracy and were all robust. Although the performance for DT was slightly lower (by 2%-3%) compared to other models. The study selected DT to be used in model development based on the reasons explained in the preceded paragraph. That closely resembles human reasoning (Kotsiantis, 2013) which makes the model easily understandable and explainable which was a priority for this study. Based on model performance, for predicting continuous variable linear regression was selected as a feature selection method and KNN for model development. While RF and DT were adopted for feature selection and model development. Table 13 indicates the type of models that were adopted as final models for features selection and model development for each decision. Table 14 shows the accuracy of the final models developed respectively for each country.

Table 13: Final models used for features selection and model development (algorithm for features selection / algorithm for model development). LM: Linear model, KNN: K-nearest neighbor, DT: Decision tree and RF: Random forest)

Prediction problem	**Ethiopia**	**Kenya**	**Tanzania**	**Uganda**
Number of exotic animals	LM/KNN	LM/KNN	LM/KNN	LM/KNN
Milk productivity	LM/KNN	LM/KNN	LM/KNN	LM/KNN
Concentrate usage	RF/DT	RF/DT	RF/DT	RF/DT

Table 14: Performance of final models developed

Prediction problem	**Ethiopia**	**Kenya**	**Tanzania**	**Uganda**
Concentrate usage (% Accuracy)	90.3%	72.0%	93%	91%
Number of exotic animals (Adjusted $R^{2)}$	0.78	0.96	0.87	0.84
Milk productivity (Adjusted $R^{2)}$	0.87	0.93	0.96	0.95

4.3.2 Model to predict concentrate usage on the farm

(i) ***Ethiopia:*** In Ethiopia, the most influential driver to determine whether a farmer would purchase concentrate was the type of feeding system adopted by a farmer (Fig. 23). Farmers who adopted other feeding systems other than grazing were more likely to use concentrate, compared to those who preferred grazing as their main feeding system. The second driver was marketing system. More than 70% of farmers who use other feeding systems apart from typical grazing and have commercial milk buyers were more likely (by 22%) to use concentrate than those with no formal markets. It was interesting to note that a farm location also determined the type of feeding system to be used. Regardless other farmers had informal markets but the fact that their farms were located in urban area

(Addis Ababa) where 20% more likely to use concentrate than those from other locations. Furthermore, the use of concentrate was associated with the best husbandry practices. Farmers with shorter lactation length (less than 15 months) were more likely to use concentrate.

Considering the other nodes, i.e. In node 6 all farmers were predicted to use concentrate and in node 19, 75% of all farmers were predicted to use concentrate. In node 6, it shows that farmers who had adopted stall feeding, had commercial buyers, their animals produced more milk (> 11.2 liters/day) and had shorter lactation length, were more likely (100%) to use concentrate. Whereas farmers in node 19 adopted grazing as their feeding system, had no specific buyers and usually liked to feed their animals' crop residue (>13 months/year). Their likelihood to use concentrate was low. Moreover, farmers who had informal markets and can easily access breeding service (walk <2.5 Km) their chances or probability to use concentrate was also high (95.4%).

(ii) ***Kenya:*** The pattern was different in Kenya, where farmers were characterized by farm characteristics and animal husbandry practices; i.e. farm production, number of hours laborer's work on the farm, feeding system, animal watering frequencies, and water sources as shown in Fig. 24. Income generated through selling of milk was considered as the main key driver, followed by watering frequencies of animals and supplementation of other animals' feed i.e. crop residue. For example, farmers who were categorized in node 5 were considered to generate low income from their dairy industry (sell ≤5 liters/day), they less watered their animals (≤2 times day), their laborer's work (≤6 hrs/day) and neither supplemented their animals with crop residue. These farmers were categorized with low probability to supplement their animals. Where on the right side of a DT, in node 33, these farmers generated income by selling ≥5 liters/day, their labors spend more time on a farm, and preferred stall feeding overgrazing. They (node 33) were classified to have high probability of supplementing (96%).

(iii) ***Tanzania:*** In Tanzania, four key drivers were identified to govern farmers' decision to use concentrate (Fig. 25). These include farm location, marketing system, animal husbandry practices, and animal productivity. The primary factor that drives farmers to use concentrate in other regions apart from Tanga was marketing system. Deworming of animals and keeping of records had an association with buying of concentrate

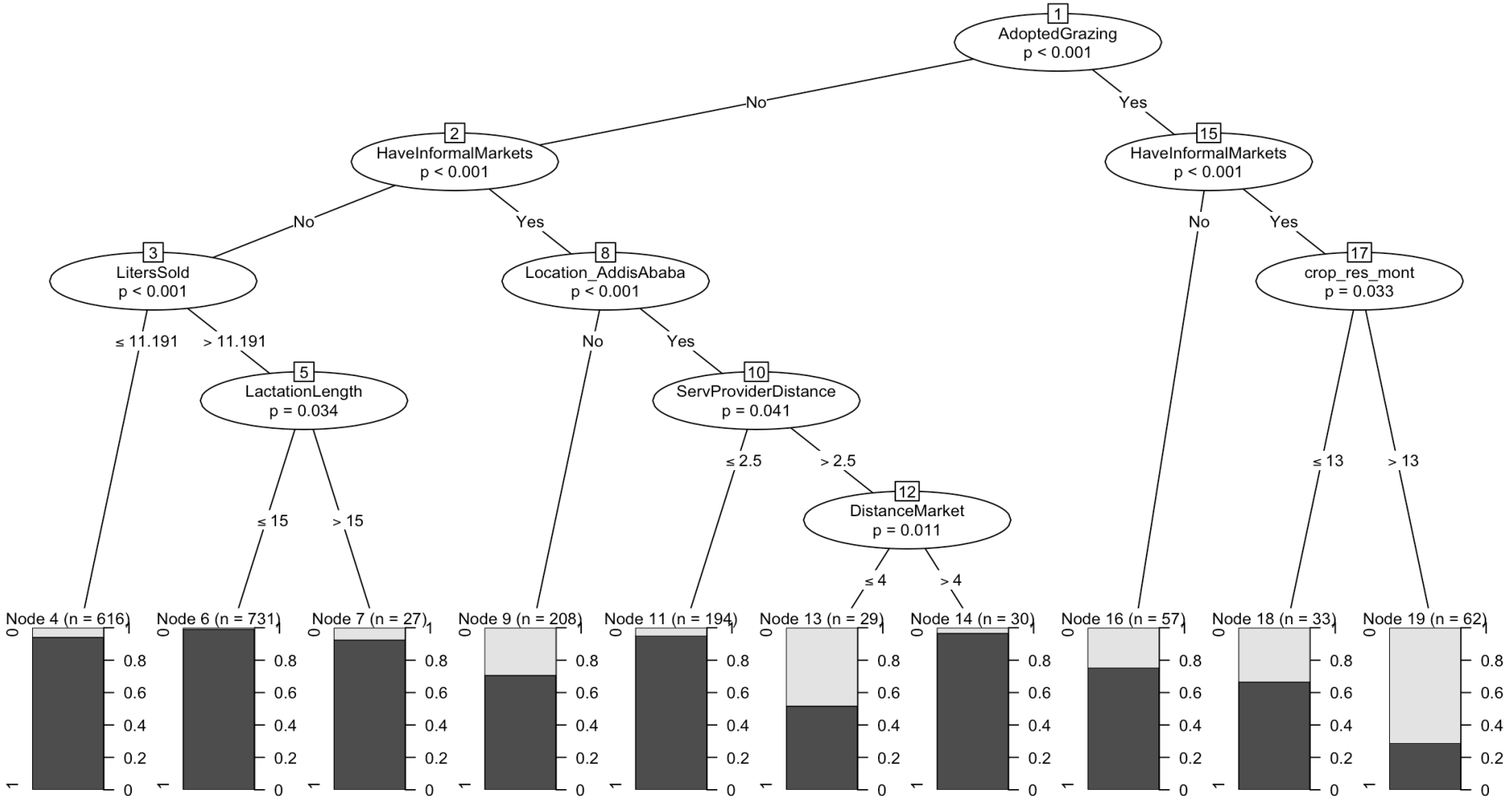


Figure 23: A decision tree model for predicting farmers decision to use concentrate in Ethiopia

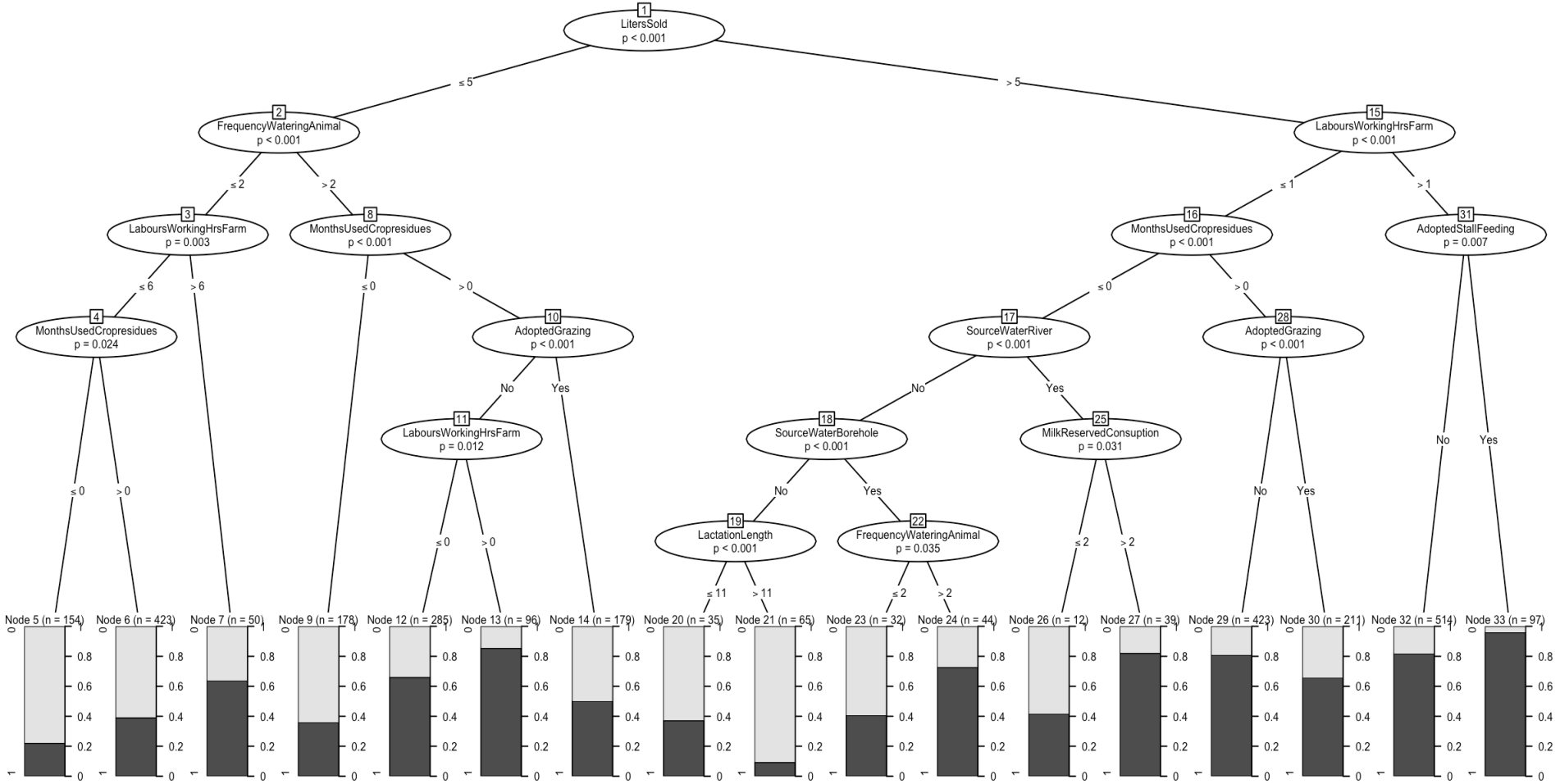


Figure 24: A decision tree model for predicting farmers decision to use concentrate in Kenya

Farmers who deworm (at least once) and those who keep records for their farm benefits had high probability of using concentrate than those who don't deworm or asked to keep records by extension officers. For example, farmers in node 6 were classified as farmers who have formal markets, their best animals produced (≥ 9.5 liters/day), keep records for their farm benefits (are not forced by extension officers) were classified to use concentrate by error percentage of 0.3%.

While those located in Tanga region (node 7), preferred to keep records for the benefits of their farms, but they sell ≤3.5 liters/day, the fact that they don't deworm their animals were classified at 100% not to supplement their animals with concentrate. However, for those (node 22) who sell milk ≥3.5 liters/day and their best animals produce >9.5 liters/day were classified to have a high probability of using concentrate. Referring to farmers classified at node 23, it was interesting to note that the fact that farmers were keeping records because were asked by extension officers decreased the probability to use concentrate

(iv) ***Uganda:*** The case for Uganda was different from that of Tanzania and slightly similar to that of Ethiopia. Where the main key driver for Uganda was the type of feeding system adopted by farmers (Fig. 26). Majority of farmers who adopted grazing as their main feeding system were less likely to supplement than those who preferred other feeding systems. Also, supplementing animals was also associated with best animal husbandry practices i.e. keeping of records. Where farmers who kept traceability and calving records were more likely to use concentrate than those who did not keep records. The study also found that farmers who adopted other feeding systems apart from grazing were more likely to use concentrate. Taking a case study of farmers classified in nodes 24 and 25. These are farmers who have adopted grazing as their feeding system, also don't trace their animals, by keeping records, neither purchase animal feeds, they also pay less for breeding service (≤70 000 UGX) and their animals produce less (≤16.5 liters/day). Their probability to use crop residue was also very low. Compared to farmers who were classified in node 17, where they prefer stall feeding overgrazing and prefer to trace their animals through keeping of records i.e. traceability and calving records and had to pay more for breeding services. They were classified to have a high probability (98.3%) of adopting concentrate.

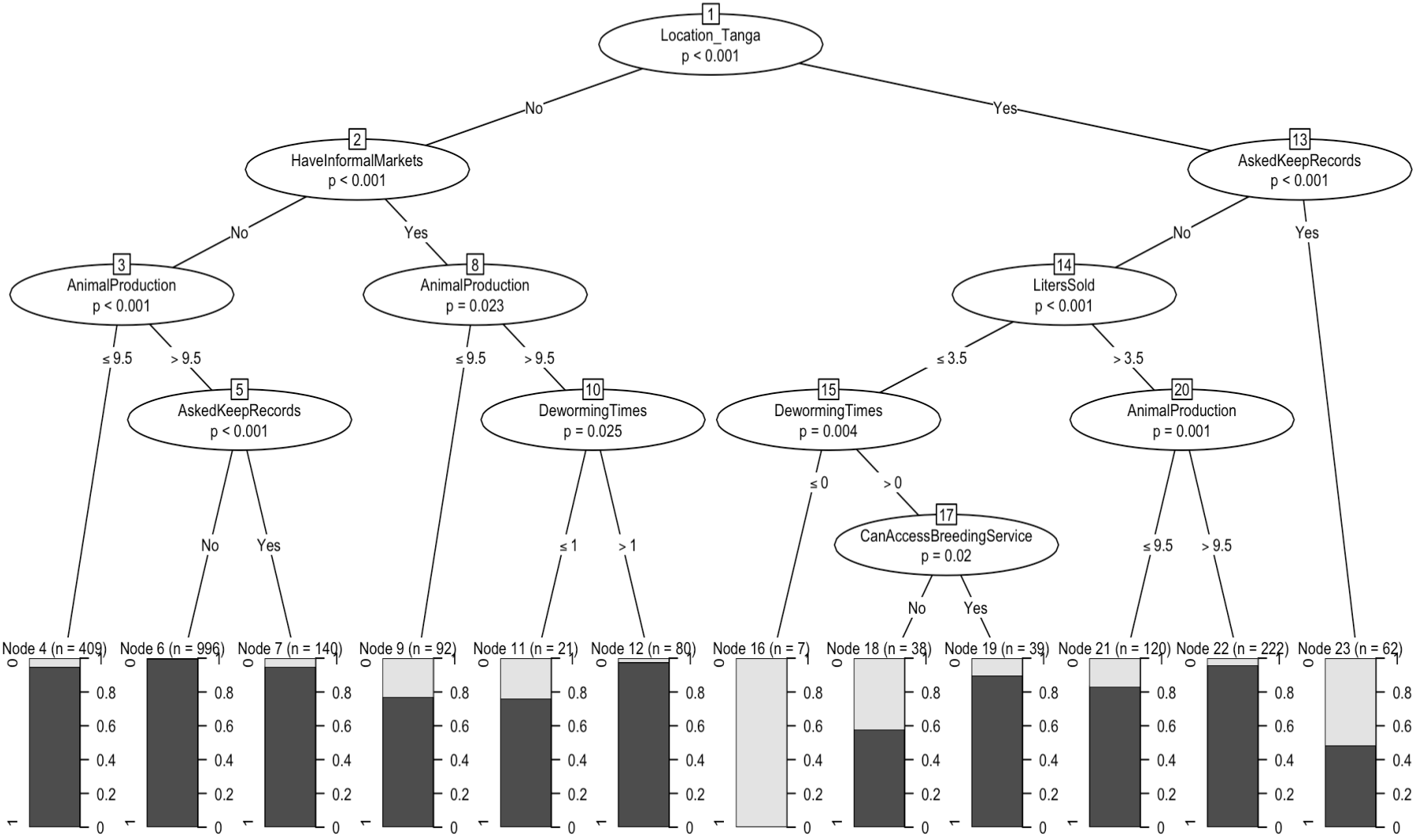


Figure 25: A decision tree model for predicting farmers decision to use concentrate in Tanzania

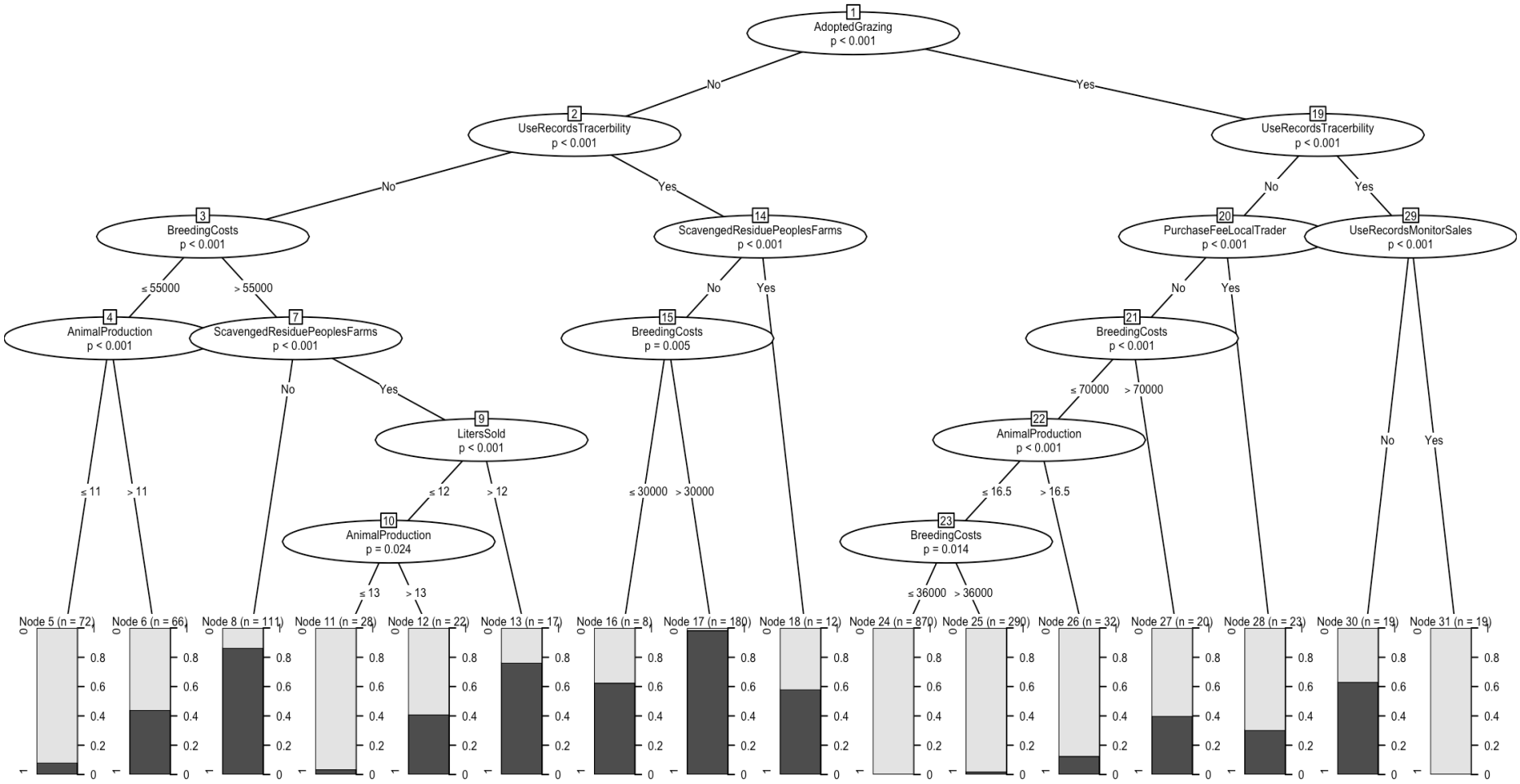


Figure 26: A decision tree model for predicting farmers decision to use concentrate in Uganda

4.3.3 K-nearest neighbors model to predict the number of exotic animals to be kept on the farm

To predict the number of exotic animals to be kept on a farm, a linear model was adopted for features selection and KNN for model development. Animal productivity was found to be highly correlated with the number of exotic animals to be kept by farmers (Table 15). It was also interesting to note that for Ethiopia the number of children a farmer had and the number of hours that laborer's work on the farm was highly correlated with the number of exotic animals to be kept on the farm.

Unlike Ethiopia, in Kenya animal identification (use of animal tags and notching) was ranked high in their significance with the number of exotic animals to be kept by a farmer. Followed by the amount of milk to be sold on a farm. Moreover, other types of animals' identification systems that were considered to be significant were the use of animals' names and markers.

The same trend was observed in Tanzania, where apart from the amount of milk produced to be highly correlated with the dependent variable the use of animal identification systems was also found to be significant in predicting the dependent variable. The use of markers, animals' names and notching were positively associated with the number of exotic animals on the farm. However, it was surprising to note that farmers from Njombe region had a negative influence on the number of exotic animals to be kept on a farm.

In Uganda, the case was different from other countries where keeping exotic animals had nothing to do with animal productivity. Instead, farm demographic information including farm location, distance to market, land usage and size, were considered to be significant predictors for the dependent variable. Moreover, farmers in Kiruhura and Mbarara had a significant positive relationship with keeping of exotic animals. While the availability of large land size which also linked with agriculture activities and growing of animal feeds shows a significant relationship to the number of exotic animals. However, farmers with informal markets (No preferred buyer or sell their milk to individual consumers) had a negative relationship with keeping of exotic animals.

All fifteen features selected by LM were used by the KNN algorithm in model development. For the KNN model high accuracy for predicting a new value of a new farm was obtained by averaging the values from K-neighbors at n-space features. Comparing the performance from cross-validation results of a training and testing model (Fig. 27 and Fig. 36), in Ethiopia and

Table 15: Variables selected by linear models to be used in developing prediction model to predict the number of exotic animals to be kept by a farmer in Ethiopia, Kenya, Tanzania and Uganda

Ethiopia			**Kenya**		
Variable	**Estimate**	**Pr(>|t|)**	**Variable**	**Estimate**	**Pr(>|t|)**
Milk sold (Litre/day)	1.45E-01	8.36E-77	Animal ID: Tags	2.00E+00	1.22E-18
Labors Working hrs/day	1.68E-01	2.69E-16	Animal ID: Notching	2.97E+00	2.75E-08
Grow grazing grasses	5.28E-01	3.32E-14	Milk sold (Litre /day)	6.11E-02	1.06E-07
Total No children	2.01E-01	5.87E-11	Distance to Extension officer	1.21E-01	1.69E-07
Provide Br services themselves	1.79E+00	1.87E-07	Milk reserved for home/day	1.73E-01	9.17E-05
Have no formal markets	1.41E+00	3.32E-07	Animal ID: Name	1.81E+00	1.35E-04
No of times used AI	3.49E-01	1.31E-05	Animal ID: Markers	5.87E+00	1.45E-04
Total Land size	1.84E-01	2.60E-05	Grow cash crops	9.38E-02	2.26E-04
Farmers Experience in dairy	1.67E-01	2.15E-04	Don’t use cattle ID	2.12E+00	2.56E-04
Farmers year at school	5.28E-02	4.06E-04	Breeding service Provided by cooperation	6.59E-01	4.18E-04
Purchase crop residue	-8.32E-01	4.40E-04	Breeding service Provided by farmer themselves	-1.23E+00	1.16E-03
Cost for water (ETH)	1.29E-02	2.36E-03	Breeding service Provided by government	1.19E+00	1.80E-03
Keep animal growth records	3.60E+00	3.01E-03	Get feeds from suppliers	1.50E+00	2.86E-03
Purchase feeds from neighbor	7.08E-01	5.18E-03	Total No children	1.00E-01	3.61E-03
No of month purchased fodder	-7.87E-02	8.47E-03	Can access to breeding services	5.50E-01	4.17E-03
Tanzania			**Uganda**		
Variable	**Estimate**	**Pr(>|t|)**	**Variable**	**Estimate**	**Pr(>|t|)**
Milk sold (Litre/day)	1.64E-01	2.10E-67	Farmers from kiruhura	1.57E+01	2.48E-149
Animal ID: markers	7.29E+00	3.66E-09	Farmers from Mbarara	6.12E+00	5.98E-40
Total No of labors	1.04E+00	1.68E-07	Total land size	1.27E-01	1.70E-26
Cost to transport milk to market	5.52E-04	1.57E-06	Farmers Experience in dairy	1.09E+00	1.57E-25
Animal ID: name	1.83E+00	5.70E-06	Grow fodder	9.05E-02	1.59E-09
Don’t use cattle ID	1.80E+00	1.17E-05	Grow food crops	5.90E-02	8.85E-07
Adopted stall feeding	5.86E+00	1.96E-05	Time to market	9.37E-01	1.49E-04
Animal ID: notching	3.88E+00	6.20E-05	Distance to market	-3.56E-01	7.38E-04
Farmers from Njombe	-1.14E+00	1.08E-04	No of month purchased concentrate	2.88E-01	1.44E-03
Best Animal production at peak	-6.34E-02	4.24E-04	Have no formal markets	-1.87E+00	2.05E-03
Farmers Experience in dairy	1.36E-01	5.44E-04	Sell milk to local consumers	-1.39E+00	2.64E-03
Asked to keep records by Extensionist	8.33E-01	7.87E-04	Farmers year at school	6.72E-02	1.84E-02
Total No children	1.13E-01	9.09E-04	Keep animal growth records	6.28E+00	2.37E-02
Service provider: farmer	1.38E+00	1.82E-03	Process ice cream	-1.86E+01	2.64E-02
Grow fodder	2.23E-01	2.48E-03	Breeding cost	1.91E-05	3.34E-02

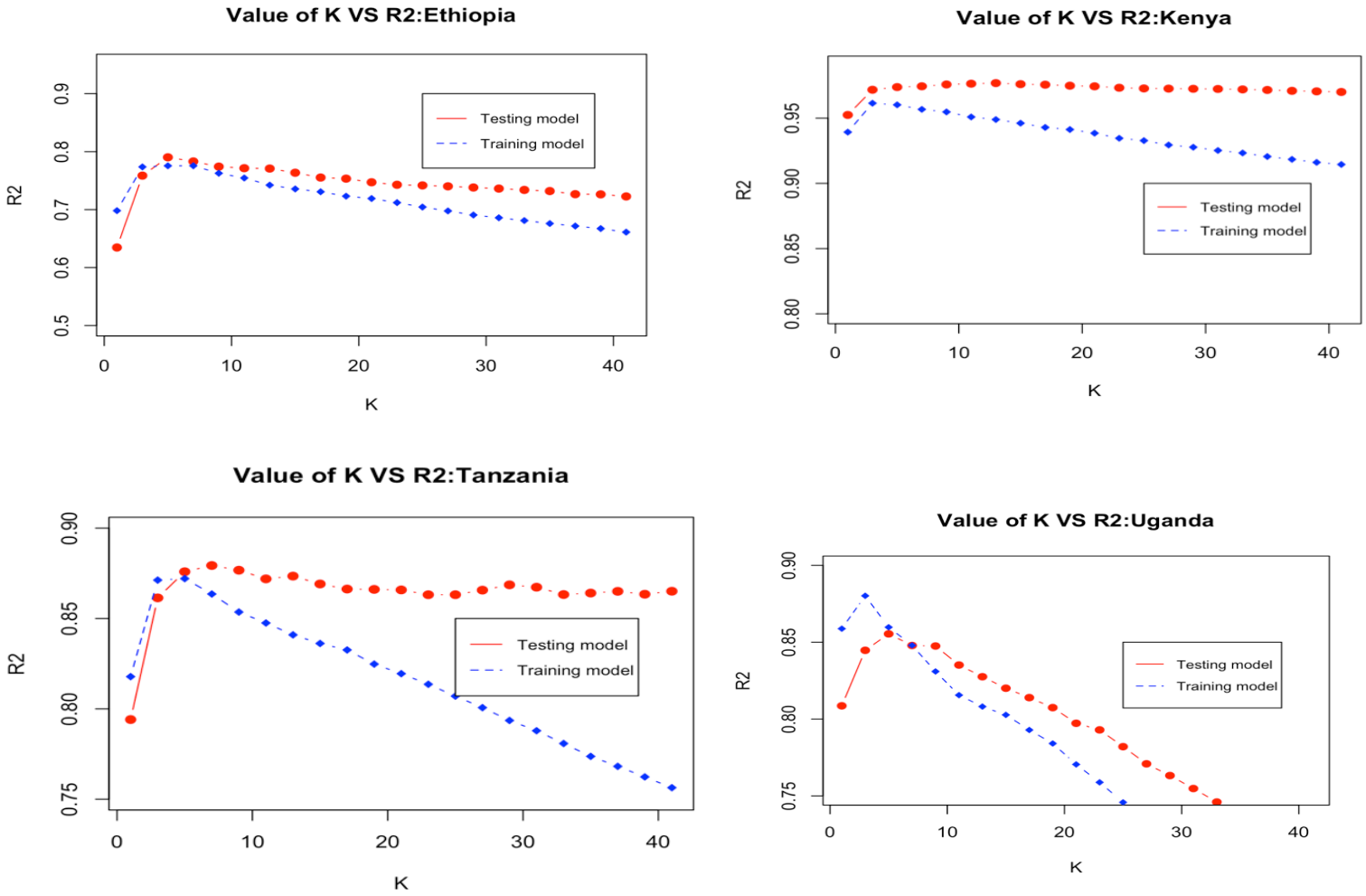


Figure 27: Displays the KNN accuracies(R^2) compared against different value of k neighbors used in predicting the number of exotic animals to be kept on a farm. Each point represents the average of ten runs of the KNN for training set and testing set

Uganda, when K=7, it yields an optimum accuracy of Ethiopia: adjusted R^2=0.78 and Uganda: adjusted R^2=0.84 for both training and testing data set as shown in Fig. 27. While in Kenya the optimum value of k=3 and Tanzania k=5 predicted the dependent variable with accuracy of Kenya: adjusted R^2=0.96 and Tanzania: adjusted R^2=0.87. However, the model performance for Ethiopia was low compared to other countries' but its performance was still considered substantial.

4.3.4 K-nearest neighbor model to predict the amount of milk to be produced on the farm

To predict the amount of milk to be produced by an animal, linear algorithm was adopted for features selection and KNN for model development.

In Ethiopia, the use of animal identification (animal names and ID tags) had a positive correlation to the amount of milk to be produced by the best animal (Table 16). While lack of formal markets (no preferred buyer or sell their milk to individual consumers) had a negative association with the amount of milk to be produced. Lactation length of animals, the frequency farmers water their animals and years a farmer spends to school also had an association with animal productivity. Additionally, farmers' location had an influence on animal productivity. For example, it was observed in this study that farmers from Asela-Shed had high milk productivity.

In Kenya, the number of months a farmer purchased concentrate significantly impacted the amount of milk to be produced by a farmer. However, it was surprising to note that the number of months a farmer used crop residue had a negative association with the amount of milk to be produced. Furthermore, the number of milking cows, the number of hours a farmer spent on the farm, the frequency farmers' water their animals had a positive association with the amount of milk to be produced. It was also interesting to note that water sources had a positive correlation with the amount of milk to be produced. Where farmers using borehole and rainwater had a positive correlation to animals' production.

In Tanzania, farm location had a significant role to play in predicting the amount of milk to be produced. Farmers from Mbeya, Njombe, Iringa and Arusha had a positive association with the amount of milk to be produced by the animal. Similarly, lack of preferred buyer or selling of milk to local consumers negatively affected milk production. While buying animal feeds

(concentrate), crop residue and the frequency of watering animals had a positive correlation to the amount of milk to be produced.

In Uganda purchasing of concentrate, frequency of watering animals and use of animal identification systems such as ID tags and keeping of records (calving records), had a positive association with the milk production. This study also noted that farmers who process their milk into other products including ghee had a positive relation to the amount of milk to be produced.

Whereas farmers with informal markets (no preferred buyer or selling milk to individual consumers) had a negative association with animal production. Nevertheless, farmers who sell their milk at dairy chilling plants had a positive correlation with milk production.

The study also investigated farmers accessibility to various services by mapping services points including breeding services, Agrovet shops, and dairy markets such as chilling plants. These were compared against a distance a farmer has to walk to the market and the average number of exotic animals per site as shown in one of a sample maps from Uganda (Fig. 28).

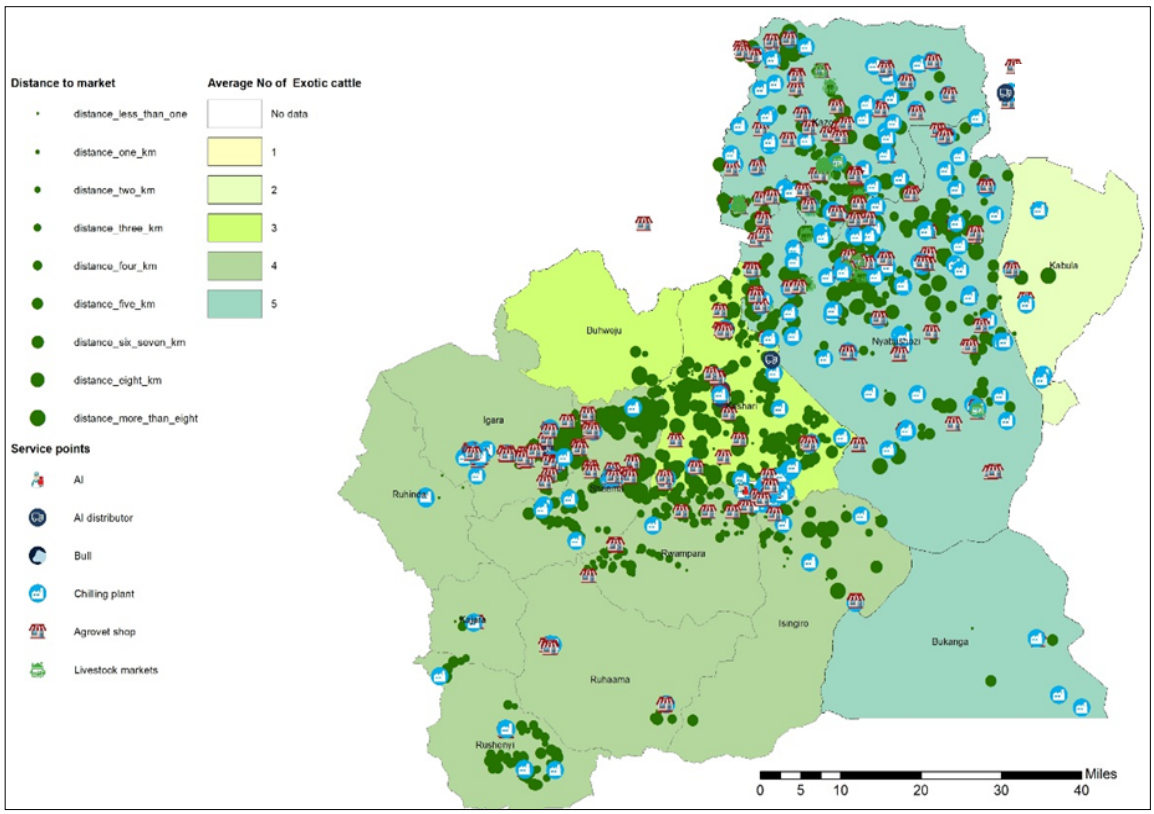


Figure 28: Farmers accessibility to various farm inputs and services including; breeding services, agrovet shops, dairy markets, chilling plant. Other information portrayed is average number of exotic animals in a given study site and distance a farmer has to work to the market

Table 16: Variables selected by linear models to be used in developing prediction model to predict amount of milk to be produced by best animal in Ethiopia, Kenya, Tanzania and Uganda

Ethiopia			**Kenya**		
Variable	**Estimate**	**Pr(>|t|)**	**Variable**	**Estimate**	**Pr(>|t|)**
Use ID name	5.20E+00	9.93E-53	No of month purchased concentrate	2.08E-01	8.11E-33
Have no formal markets	-3.98E+00	1.22E-34	No of milking cows	6.24E-01	1.93E-24
Lactation length	2.79E-01	2.17E-28	No of month purchased crop residue	-2.00E-01	2.26E-14
Don't use cattle ID	4.46E+00	2.15E-22	Labors Working hrs/day	1.63E-01	1.79E-13
Animal ID: tags	2.34E+00	1.97E-16	Frequency of watering animals	5.41E-01	3.24E-11
Sell milk to local consumers	-2.12E+00	2.58E-09	Cost for transporting milk to market	6.54E-03	1.76E-06
Farmers Experience in dairy	1.07E-01	5.21E-09	Source of water: borehole	7.73E-01	1.41E-05
Frequency of watering animals	6.66E-01	2.21E-07	Source of water: rain	8.25E-01	2.61E-04
Farmers from Asela Shed	2.95E+00	2.65E-07	Market: Dairy chilling plant	7.63E-01	3.86E-04
No of milking cows	3.69E-01	3.75E-04	Distance to buyer	5.99E-02	4.71E-04
Time taken to market	-6.85E-01	6.58E-04	Use crop residue from their farms	7.18E-01	7.01E-04
Grow grazing grasses	2.90E-01	7.40E-04	Farmers from Central	5.86E-01	2.65E-03
Breeding method recently used	8.70E-01	1.01E-03	spend	4.14E-03	4.65E-03
Number of times used AI	2.93E-01	2.92E-03	Market: Milk collection center	1.58E+00	5.00E-03
Income from crops	1.02E-07	4.61E-03	Why dairy cooperative	-2.25E+00	1.31E-02
Tanzania			**Uganda**		
Variable	**Estimate**	**Pr(>|t|)**	**Variable**	**Estimate**	**Pr(>|t|)**
Farmers from Mbeya	4.65E+00	3.67E-61	No of month purchased concentrate	2.37E-01	7.42E-11
Farmers from Njombe	2.82E+00	1.55E-18	Have no formal markets	-1.48E+00	7.87E-11
Have no formal markets	-2.04E+00	3.14E-11	Frequency of watering animals	5.34E-01	4.75E-10
No of milking cows	7.92E-01	4.04E-10	Sell milk to local consumers	-1.04E+00	1.71E-08
No of month purchased concentrate	1.15E-01	4.92E-08	Process ice ghee	2.34E+00	5.21E-06
Lactation length	1.92E-01	9.29E-08	No of milking cows	1.79E-01	8.26E-06
Market: Individual consumers	-1.09E+00	4.29E-07	Lactation length	1.17E-01	2.49E-05
Farmers from Iringa	1.53E+00	2.05E-06	Keep animal calving records	1.71E+00	1.79E-04
Purchased crop residue	2.08E+00	5.90E-06	Market: Dairy chilling plant	7.58E-01	2.65E-03
Frequency of watering animals	3.62E-01	6.89E-06	No of local animals local	-1.31E-01	2.93E-03
Breeding service Provided by government	-1.90E+00	1.30E-05	Use records for animal identity	1.14E+00	4.39E-03
Farmers from Arusha	1.43E+00	1.76E-05	Animal ID: tags	6.44E-01	1.07E-02
Animal ID: name	1.85E+00	2.53E-05	Keep animal growth records	-2.72E+00	1.53E-02
Grow animals' fodders	5.95E-01	3.81E-04	Cost to transport milk to buyer	3.17E-04	1.62E-02
No of times used AI	3.96E-01	4.06E-04	No of times used AI	2.63E-01	1.65E-02

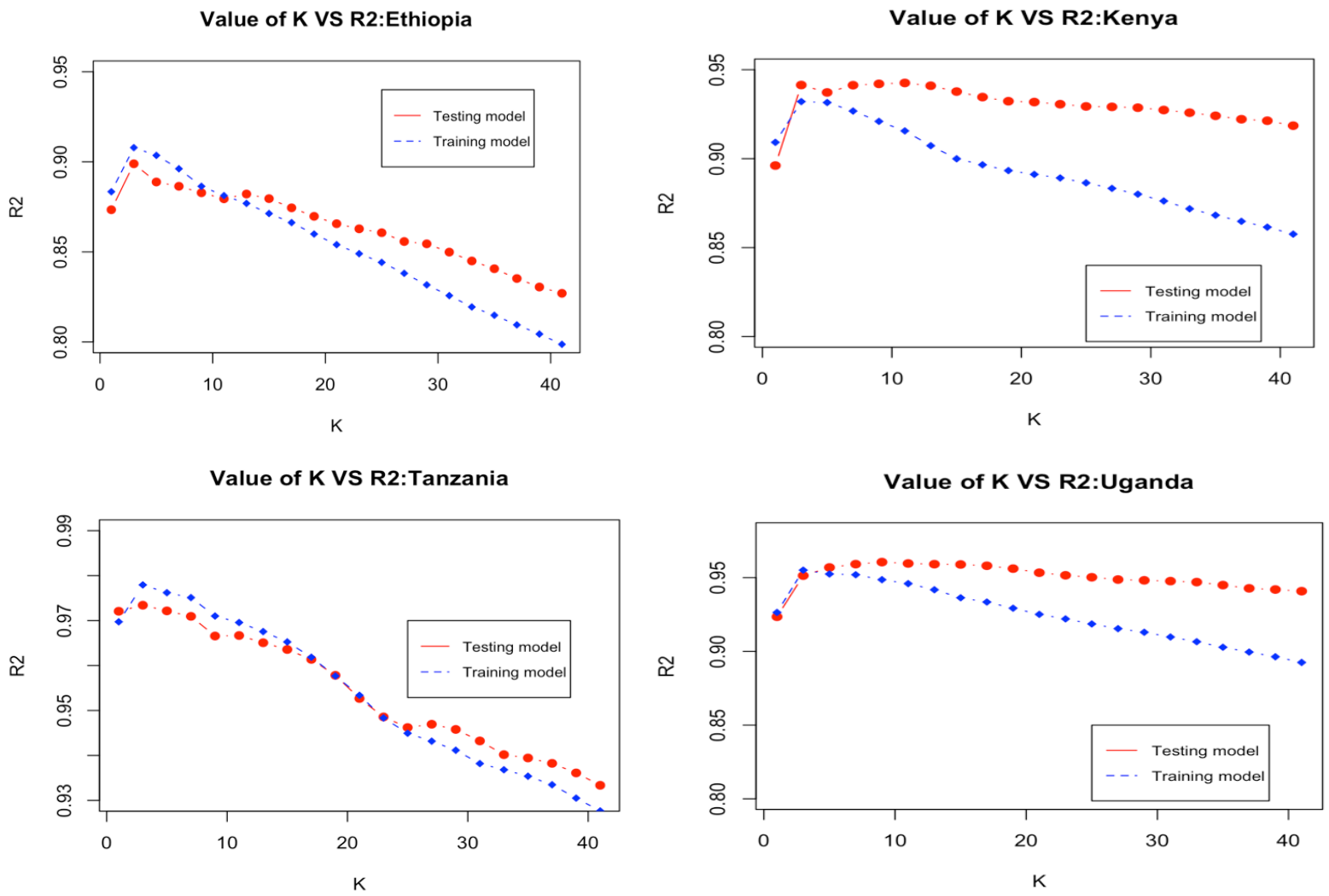


Figure 29: Displays the KNN accuracies(R^2) compared against different value of k neighbors used in predicting the amount of milk to be produced on a farm. Each point represents the average of ten runs of the KNN for training set and testing set.

It was observed that a number of chilling plants which were mostly saturated to regions with denser dairy animal population (exotic animal). This pattern was found in Uganda and Kenya. Ethiopia and Tanzania had few chilling plants and scattered unlike, Uganda and Kenya.

All fifteen features selected by LM were used by the KNN algorithm in model development. An optimum accuracy was obtained when the value of K=11 in Ethiopia with the prediction accuracy of adjusted R^2=0.875. In Kenya k=5 which yields a prediction accuracy of adjusted R^2=0.93 and Tanzania the value of k=17 with prediction accuracy of adjusted R^2=0.96. While for Uganda k=3 and maintained a prediction accuracy of adjusted R^2=0.95 (Fig. 29 and Fig. 37).

4.4 Models validation

The validation process was performed using Rwanda data as elaborated in section 3.5. Since all decisions to be modeled were qualitative variables a combination of DT and RF was used for modeling. Generally, all models attained high accuracy as shown in Table 17. In predicting whether a farmer will supplement their animals or not the model attained an accuracy of 94.2%, whereas in predicting the type of breeding method to be used by a farmer the model attained an accuracy of 82%. In predicting whether a farmer will continue keeping a DIRINKA animal the model predicted at the accuracy of 70%. The model performance attained proved that the algorithm used for modeling in this study were robust enough even to a new set of data.

Table 17: Model evaluation using Rwanda data

Farmers Decision	Features selection Method	Model Development	Model Accuracy
Use of supplements	Random forest	Decision trees	94.2%
Breeding method (AI or Bull)			82%
Continue keeping a DIRINKA animal			70%

The results obtained show that a farmer's decision to supplement was influenced by their decision to purchase other types of animal feeds (Fig. 30). Farmers who were not purchasing animal feeds their probability of supplements was very low. Also, the type of cattle kept, and education level had implications on farmers' decisions to supplement. Farmers with advance or university level education were more likely to supplement than farmers with low education

levels. Moreover, farmers who preferred to keep exotic breeds such as Friesian and Jersey were more likely to supplement than those with local and crossbreeds.

It was also interesting to note the influence of farmers' training in decision making (Fig. 31). Whereby it was established that farmers who received training on best animal husbandry practices had a high probability of using AI (i.e. node 18). Compared to other groups of farmers who were not trained (node 8).

Moreover, the main key drivers that determined whether a farmer will continue keeping the animal project was the orientation of the farm which can be linked with available resources, animal productivity i.e. if the animal given had calved and the person who was responsible for taking care of the farm (Fig. 32). Taking a case study of farmers who were classified in nodes 3,4 and 22; Farmers in node 3 were from Amajyaruguru and their animals that were given by the project were yet to calve. While those in node 4 their animals were calved which slightly increases the probability for them to continue keeping the animals. For farmers in node 22 were located at Amajyepfo and the people who were engaged in taking care of animals are members of the family except the husband and they were growing cassava, so had other sources of food and income. They were feeding their animals more (>55kg/day), their probability to continue keeping the program animals was high.

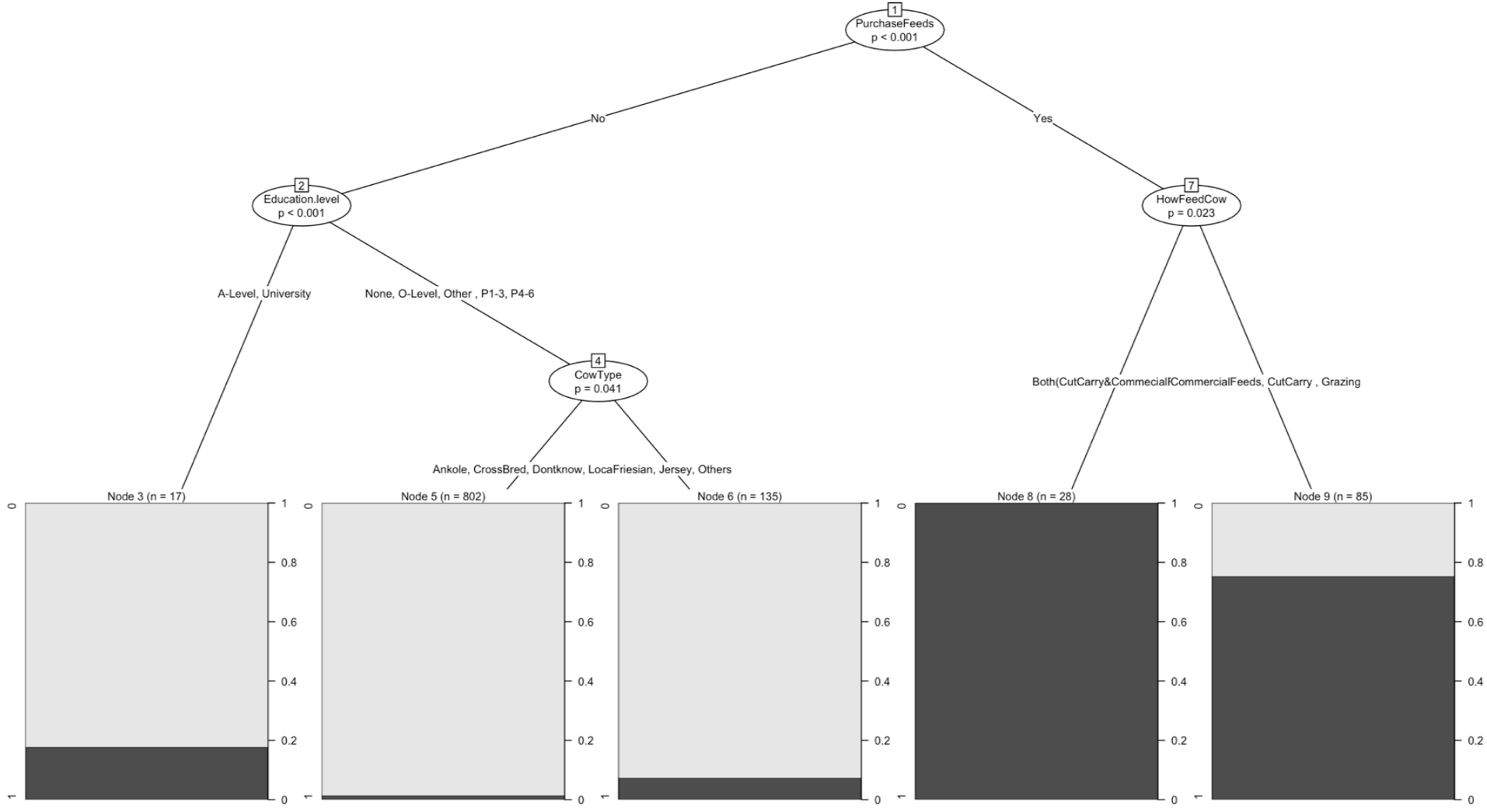


Figure 30: A decision tree model for predicting the use of animal supplement in Rwanda

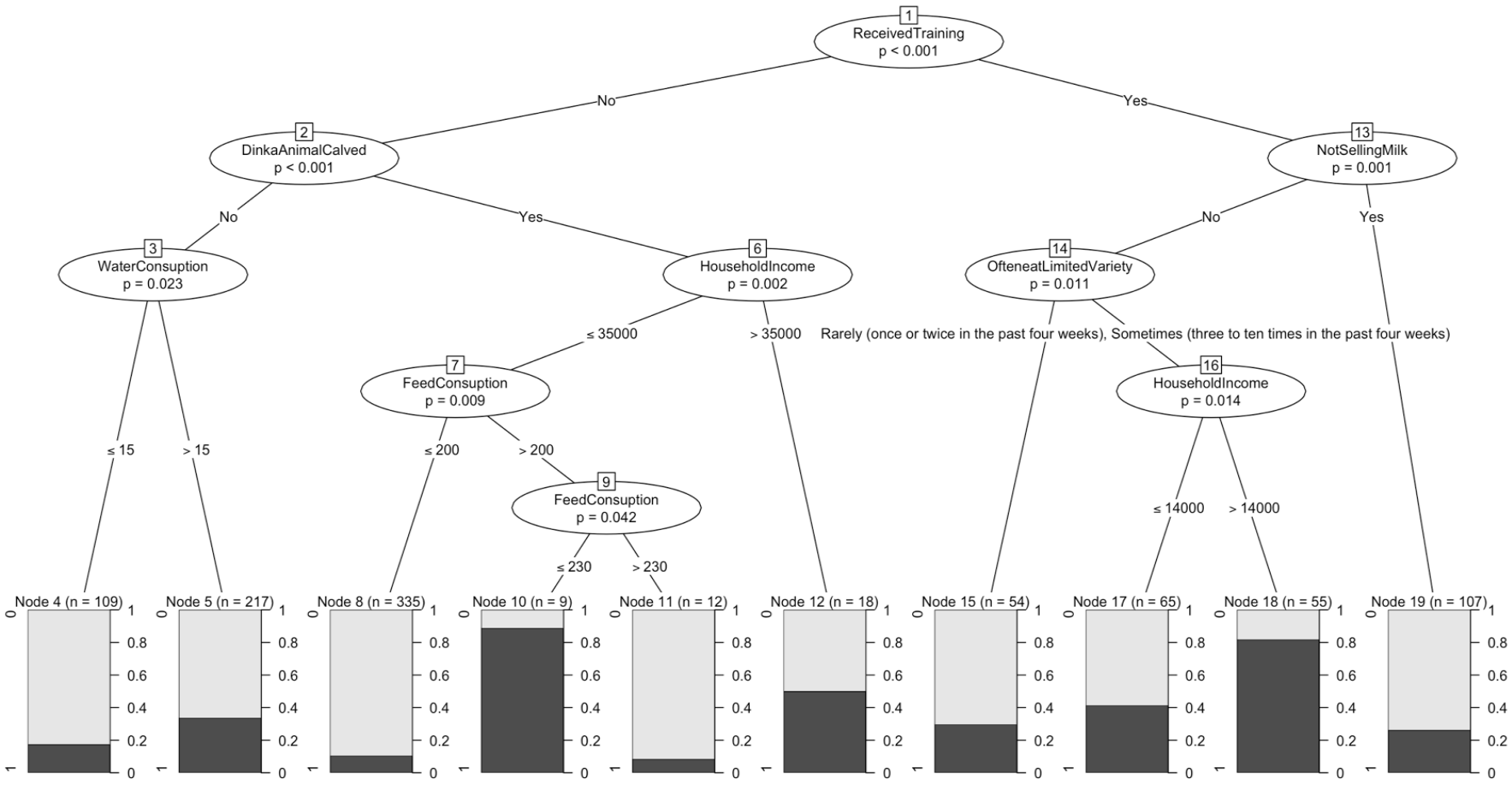


Figure 31: A decision tree model for predicting adoption of AI as breeding method to be used on the farm in Rwanda

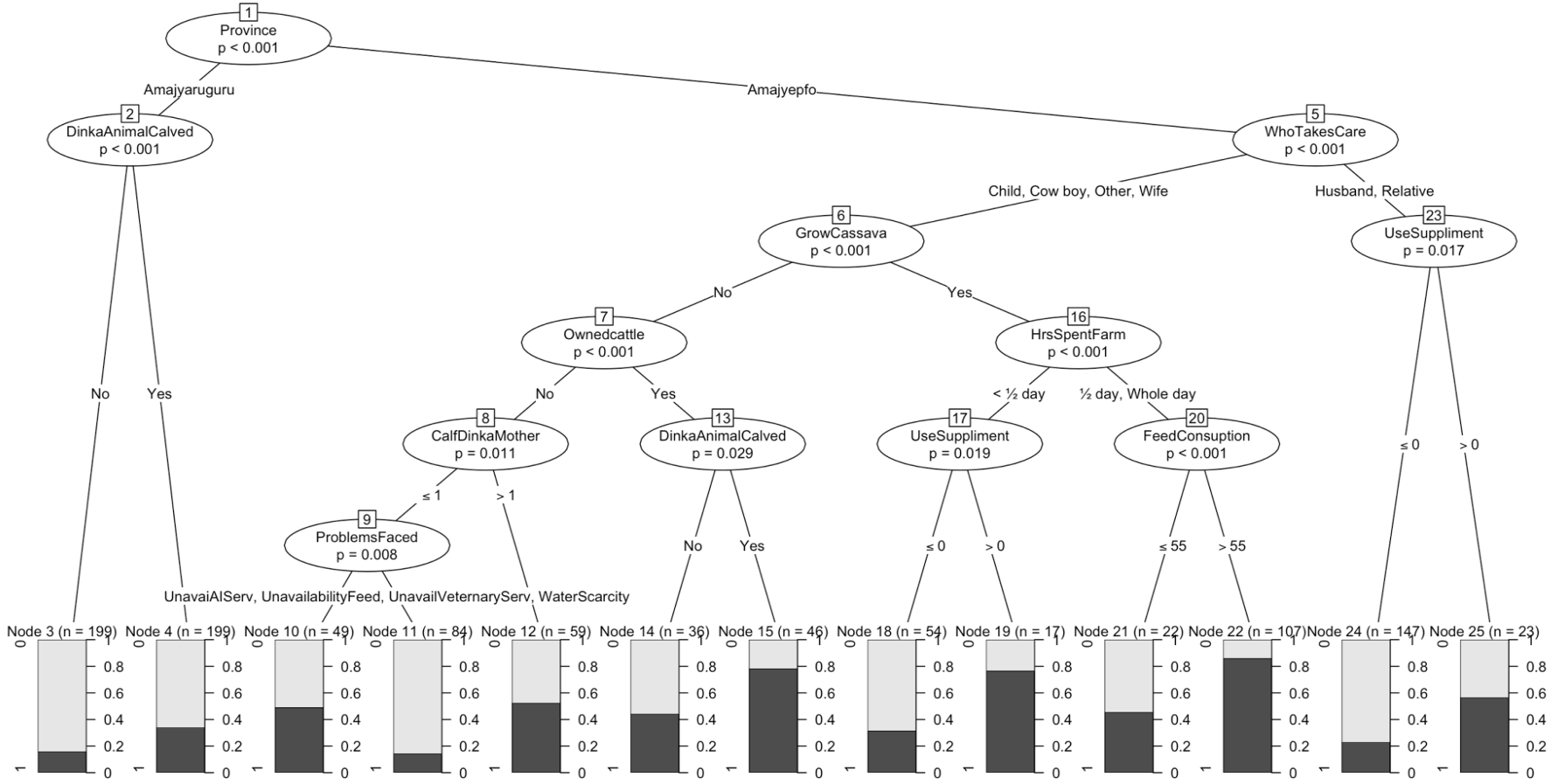


Figure 32: A decision tree model for predicting whether a farmer will continue to keep a project animal (DIRINKA) in Rwanda

4.5 Discussion

The government is setting various strategies; from policies and other supporting initiatives to support farmers attaining maximum productivity, including providing farmers with various services, i.e. breeding, health, feeding to mention but a few. In this regard, various technologies to improve productivity has been established or proposed for the farmer to adopt. However, there has been low adoption of these technologies and most of the time reasons that drive this pattern remain to be unknown at a farmer's perspective (Kabunga, 2014). Moreover, in developing policies, setting up the budget and allocating resources is always vital to know farmers' demands and preferences. Therefore, this study responds to some of the questions which are being asked by decision-makers on factors that influence/hinder farmers to adopt technologies/ or abide by best husbandry practices. However, identifying factors alone is not as informative as projecting and forecasting, which is known to influence more evidence-based decision making and resource management. This study also aimed at developing models that can be used to identify factors that influence farmers' decisions but also being able to predict decisions to be made by a farmer from a given set of factors. As a result, four models have been developed, namely (a) model to predict the adoption of AI as a breeding method on the farm (b) model to predict usage of animal supplements (concentrate) (c) model to predict farmers' decisions to keep exotic animals and (d) model to predict the amount of milk to be produced on a farm.

The differences in feature selection methods signify that each algorithm has its own preferences. That can be due to various reasons including data types (Numeric, Decimals, etc.) and a relationship between the dependent and independent variables (Singh, Halgamuge & Lakshmiganthan, 2017). It was also noted by Fawcett (2015) that not all patterns found via data mining are "interesting." For the data patterns to be "interesting," they should be logical and actionable. Therefore, at some point, it requires a human intervention to extract knowledge (Ristoski & Paulheim, 2016). Also, since no single selection technique is capable to fully forecast which variables will be effective in another modeling tool, an intensive searching process needs to be involved. Sometimes applying the same algorithm to slightly different data produces a very different model. For example, in this study, the neural network and GMM produced very different prediction accuracy depending on which selected feature sets were used for training. Also, it was expected that the Neural Network would perform better than Random Forest but that was not the case for this study.

Based on the test performance of the models, all models attained higher accuracy, which demonstrates that despite the complexity of the dairy sector, still, ML was able to capture features and model farmers' decisions efficiently. The good prediction accuracy attained by classification models signifies that all features selection and model development algorithms were useful in their own rights. However, for each feature selection method to have its own preferences within the same set of data, it indicates that algorithms are not the same. Hence, intensive searching for the appropriate features and understanding the domain science is also vital. Despite models' performance, it was also important for this study to tradeoff along with various aspects in selecting the appropriate model to be adopted for model development.

This study adopted a combination of RF and DT for features selection and model development. Random forest was chosen because of its robustness but also because it gives a wide range of variables. Random forest has been widely used in different fields to improve the accuracy of predictive models (Shaikhina *et al*., 2017). In dairy production RF models have been used to predict profitability ratio of dairy farms based on financial and production-related variables, animal productivity such as prediction of conception success and dystocia detection (Hempstalk, McParland & Berry, 2015b; Singh *et al*., 2017; Zaborski *et al*., 2017). Also, the use of nonlinear models such as DT, RF, KNN, GMM and NN has an advantage over linear models as are able to map highly non-linear heterogeneous input and output patterns even when physiological relationships between model variables could not be determined due to complexity (Shaikhina *et al*., 2017). Which is a common case for the dairy sector as it was observed that one decision can be influenced by various factors that are nonlinear.

In addition, using a DT algorithm to define classification rules for modeling farmers' decisions led this study in identifying very useful findings/patterns that could not be identified by other algorithms. In this study, DT provided a transparent method for inductive learning of data (Lior, 2014). When nodes split, they provide useful information on what specific level of statistical significance different factors were associated with adoption or decline of service/technology. The decision tree provided a simulation tool that was able to classify and explore farmers in various groups based on their scenarios/characteristics. Such information assists a decision-maker to define the relationship among factors but also to identify a targeted group for intervention. However, DT at some point is considered to be unstable algorithms as it defines its classification/regression rules based on the size of a training set (Shaikhina *et al*., 2017). But this study used a significant number of the training set and the testing results revealed that

the models were robust. Also, it was possible to aggregate many different trees and averaging across them which substantially helped to improve the performance.

The model's performances for a DT, RF, and LR was not the same for regression problems (predicting the number of exotic animals and animals' production). This can be due to the reasons that DT and RF performed better with classification rather than regression though they claimed to do both. It has been argued that these are some cases where linear models can be better than RF or DT and this is when the underlying function is truly linear and when there are a very large number of features especially with very low signal to noise ratio (Shaikhina *et al*., 2017); a case when such models (RF and DT) can have a little trouble to model a linear combination of a large number of features.

The results of this study offer a number of practical implications for the dairy industry. They give insights to decision-makers including policymakers but also breeding units, farm inputs suppliers and services providers. Here are some of the important findings.

(i) The type of feeding system adopted by a farmer to drive farmers' decisions to supplement their animals with concentrate highlight several facts. Where farmers who adopted grazing are less likely to supplement and this is due to the reason that animals spend more time during grazing hence lack time to supplement and vice versa was true. In this study, the case was predominant to farmers in Uganda where more farmers preferred grazing over stall feeding. This pattern can be linked to land accessibility. Farmers in Uganda owns a bigger land size; two to three times of total land owned by other farmers in Eastern Africa countries (Mwanga *et al*., 2018). This allows farmers to grow animal feeds. Hence, concentrate demand for farmers in Uganda will be low as farmers choose to supplement with fodder.

(ii) Similarly, adoption of feeding systems can be linked with their agricultural activities as it was observed in Ethiopia; where farmers supplement their animals with crop residues harvested from their farms. In this respect, it can be generalized that farmers who have fully reliable access to animal feeds such as fodders or crop residues are less likely to use concentrates. However, this raises a concern on-farm management practices of small-scale farmers. Whether those who adopt have prepared to incur a lot of costs in purchasing animal supplements and vice versa is true. With this regard one of the solutions to these clusters of farmers is to assure that farmers feed good quality grass or concentrate (with production

nutrient requirements) and if that is not sufficient any supplemental interventions or strategies to be implemented on a farm, feed suppliers should target this group of farmers.

(iii) It was interesting to find that the presence of formal markets had a positive association with milk productivity which was also true to the keeping of exotic animals and supplementing animals with concentrate. Marketing is considered as a driving factor for dairy intensification (Lemma, Mengistu, Kuma & Kuma, 2018). This is a common case to any business where markets define objectives of the producers. Therefore, relating this to a dairy perspective where farmer's objectives towards a dairy matter (having a business perspective). Because it was identified in this study that, even those farmers who were specified to sell their milk to local buyers, which is still informal markets had a low probability of supplementing. This scenario justifies the fact that it is not only about having a milk buyer but having a formal market such as chilling plants, milk ATM, processors, etc. As it was observed in Kenya and Uganda where these two countries had a number of dairy chilling or milk collection centers (Fig. 28). This also gives farmers the reasons to incur more costs to supplement because in turn they are guaranteed payoff. The scenario can be linked to the pattern that was identified in Kenya where a farm production was considered as the top significant driver to determine farmers' decisions to supplement.

(iv) In addition, the existence of these plants has added more value to farmers. Farmers have been receiving support such as training and other services i.e. breeding services. Moreover, the system has been diversifying the income source of a farm by providing farmers with a loan and pay through milk that they will be provided to the plants. This signifies that dairy production in Eastern Africa can be made more sustainable by supporting farmers with marketing issues.

(v) However, for small scale farmers, a smart marketing approach needs to consider the farming system. It was established that market orientation can also be defined by farming systems (rural and Urban or peri-urban dairy systems). Farmers located in rural areas have no direct access to local consumers market (Satterthwaite, McGranahan & Tacoli, 2010). As a result, they rely on milk collecting centers or middlemen. While those in urban means their major part of the milk is marketed through the informal market which at present is often more profitable than formal markets. Although, for this study it was found that access to formal markets had more influence in productivity and vice versa was true. This can be due to the fact that regardless of farmers have access to informal markets still it is not

sufficient for their needs, which was the opposite to those who had formal markets such as chilling plants and milk collection centers. Therefore, governments can strategize in establishing milk collection centers or processing units especially in rural areas and some parts of urban areas where the local consumer market is not flooded to ensure a reliable market to farmers.

(vi) The model predicted that farmers who were located in Tanga, Tanzania were associated with adoption of best husbandry practices. This can be related to various factors. Apart from farmers depending mainly on dairy as their economic activity it has being a well-established and a major dairy zone in Tanzania. Most farmers in this region have a reliable market where they sell their produce to Tanga Fresh, which is the main dairy plant in Tanzania (Nell, Schiere & Bol, 2014). Moreover, these farmers are members of Tanga Dairy Cooperative Union (TDCU) which empower farmers and strengthening their dairy farms through training, giving loans etc. Despite the current achievements, several technical advisory missions have been conducted via farmer associations by NGO and private investors. Thus, for efficient and sustainable dairy production, the same initiatives can be extended across regions. This scenario also highlights that during policy development there is a need for conducting location scenario analysis. This will help to explore different options, such as preexisting infrastructures for proper allocation of resources.

(vii) Based on the results of this study, it's obvious that in Ethiopia and Tanzania the task for providing breeding services to farmers appeared to be shared between the government and private practitioners. This can be attributed to the policy shift caused by reduction of government involvement in the provision of this service. Which it has been a major concern that most private AI services are underdeveloped (Mugisha *et al*., 2014). Most of these private sources work in poor environment i.e. Lack transport and sometimes can be running out of nitrogen which has implications to the quality of semen. Since most of them don't want to make a loss they have been serving farmers without considering that the semen may have expired. As a consequence, farmers consider the costs of AI services as high leading to the number of repeats that have undergo for a successful conception rate as it was observed in this study. In order to succeed in this policy shift, the government need to assess the efficiency of private sectors, promote their expansion and help to strengthen their business. Moreover, it was emphasized that the involvement of the

government in providing breeding services to farmers should not be ended. Thus, there is a need for it to continue subsidizing the service.

(viii) Following up on breeding service it was established that farmers preferred to use bulls over AI due to unavailability of AI service and its weakness, i.e. more expensive and had many repeats. Due to the reason that most farmers could not afford to keep a bull on their farms due to the small size of land and maintenance costs as it was observed in Ethiopia, Kenya and Tanzania, farmers rely on neighbors. Most of the time this service is given for free. Thus, if AI technology has to be adopted one of the initiatives to be taken by service providers is to guarantee that the service is available. While on the other hand, there is a need to resolve on smooth transformation and ensure the service is reliable i.e. efficiency of the services. Additionally, other strategies that can be taken by breeding units and service providers is to raise awareness and promote the service to farmers through training and pilot studies (Dadzie, Amponsah, Dadzie & Winston, 2017). As it was observed in Rwanda that farmers' training had a significant inference for farmers to adopt AI.

(ix) It can be generalized that most of the decisions made by farmers anchored around the type of reproductive system adopted on the farm. Farmers who were surveyed in this study can be divided into two major reproductive systems; rural and Urban/peri-urban smallholder dairy. Where rural farmers, mostly are mixed farmers and own large land size that can opt to keep their cattle under semi-zero grazing systems. Meaning they can have option of grazing but also feed their animals from cultivated fodder, crop residues and cut grasses from waste or communal land. These farmers often are limited to local consumers market left with no option rather than relying on formal markets such as chilling plants or milk collection centers. Compared to urban/peri-urban who mostly have limited access to land which in return has to spend more to purchase animal feeds. Also, their major part of the milk is marketed through the informal market. It is obvious that the two groups will have different preferences on their farm inputs. Hence one can appreciate the need of having the right mix of policies that can cover different segments /classes for farmers.

(x) Lastly, the study shows that decision-makers can consider dividing farmers more into other two groups, intensive and traditional dairy farmers. The traditional farmers are those who adopt practices that need no investments, i.e., no record-keeping, or deworming or purchasing any type of animal feeds. Their practices are also reflected on their productivity

because they have low milk production. Thus, to continue improving dairy productivity can focus on intensive farmers while continuing emphasizing tradition farmers to adopt best husbandry practices. Also based on different farm dynamics in different countries that were found in this study, any policy to be developed needs to be investigated along the regions and try to cluster farmers for evaluating if the developed policy would work for all clusters of farmers.

CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

The focus of this research was to find and assess the potential areas where machine learning (ML) can be applied in the livestock sector specifically in the dairy industry. In chapter 1 we briefly describe the need and why the use of advanced technology such as ML is now important in enhancing policymaking, planning and in strategic decision-making. Our analysis found that the dairy sector is challenged with a number of problems. One of the major problems that have been addressed in this study is the lack of analytical decision supporting tools and mechanisms to identify the constraints to livestock production, analyzing farmers' demands, preferences and factors that hinder them from adopting various technologies that are essential in improving productivity. Mostly the policy priorities and directions for service delivery get determined without enough supporting evidence. This has resulted in the failure of many projects, strategies, and plans which aim in improving dairy.

Given the challenge above our main objective was to demonstrate the potential of ML technology in enhancing evidence-based decision making, by developing models that can be used in analyzing farmers' demands, preferences and identifying factors that constrain their choices. As part of our objectives, we modeled three decisions that a dairy farmer makes daily including farmers' decisions in regard to breeding methods to be used on the farm, use of concentrate as part of their feeding system and keeping of exotic animals. Furthermore, we modeled the implication of these decisions made by farmers and other factors in animal productivity.

At the outset of this work, we first strived to understand how farmers make decisions and exploring various factors that affect their decisions. This helped us to produce general knowledge on how the models should be designed and come up with a list of various factors that were hypothesized to influence farmers' decisions. Then we deployed ML techniques which were used to draw various patterns of information in regard to decision making. Further, we evaluated the robustness of these models by testing their performance with a new set of data.

To address the first objective, we selected one decision; farmer's decision to select a particular breeding method into gain an insight on how farmers make such decisions and how different factors can influence their decisions. For this, there were several factors that influenced farmers to make certain decisions. These include farmer and farm characteristics, institution settings, farm income and costs incurred to maintain a farm. In this regard, we experienced that learning from data can assist in the discovery phase of different patterns from the data. Also, we observed that it can be very complex to model farmer's decisions as one decision can either be constrained or influenced by several factors. Meaning that model development requires an intensive searching of features. We also observed that there were several dynamics among the regions, that one solution can not fit them all. Hence each country may require its own model.

The second objective was to demonstrate how can ML be used to provide insightful information to decision-makers. In developing models, we started by identifying key predictors from a pool of more than 120 variables. We begin by modeling one decision (Use of AI) into get an insight about our data and how to go about with other decisions. Six models were compared. These were LR, NN, DT, KNN, RF, and GMM. Two models (NN and GMM) were dropped due to their poor performance in some data sets. Hence in the second task of modeling other decisions; keeping of exotic animals, supplementing animals and animal productivity, only four models were considered against all three features selection algorithms. Then we examined the performance of each model. A combination of RF and DT was used for classification problems. The results of this study proved how advanced technologies such as ML have the potential of improving decision making to the smallholder dairy sector.

The third objective was to test the models' efficiency and robustness. Data from smallholder dairy farms in Rwanda were used for the validation process. All models that were tested also attained high accuracy. Some of the contributions of this study are described below.

We have analyzed more than 40 papers featuring the application of ML in addressing dairy problems that provide relevant insights to researchers. We have studied and examined various applications of ML in dairy. In each of these studies we looked into the problems that have been addressed, a solution proposed, ML algorithms used, and nature of the study including study site. Also, we investigated the types of data used and if possible, data collection devices that were employed in the implementation of the study. Through this analysis we have realized that ML has the potential of improving dairy productivity; including automation of various

processes on the farm which could be useful to farmers as they can be able to monitor their farms from a distance.

Similarly, despite the potential of these tools in facilitating decision making we observed that the use of ML in enhancing decision making to other livestock stakeholders is still in its infancy stage. Our aim here has also been both to encourage researchers to build upon previous research to a more widespread realization of ML and the implementation of these tools to small scale farmers in addressing various challenges that are faced by farmers: such as early detection systems for diseases, estrous and allow consistency monitoring of a farm. Moreover, the governments can assist farmers by using facilities such as Agriculture Development Bank in collaboration with Ministry of Agriculture to train farmers on the importance of these tools and provide these tools on loan basis or by subsidizing the cost of these technologies

Also, in this study, we have shown that ML increases the confidence in making decisions. Country-specific predictive models that were developed in this study can be used in predicting farmers' demand in regard to farm services. Furthermore, ML can be used to extract factors that influence their adoption and preferences. The models presented in this study will be helpful to decision-makers including policymakers, services providers and other livestock stakeholders in creating policies, improving farm services by making informed advice, but also predicting farmers' demand for proper allocation of resources.

Moreover, the results of this study can be used as a hands-on reference by different organs of the livestock sector including policymakers but also services providers such as breeding units, livestock extensions', researchers, and inputs suppliers. The models give an overview of the key drivers that govern the sector. For example, the findings that show farmers will opt to use AI service if they have ever used it before or have recently used it. It gives insight to the decision-maker that one of the initiatives to be taken is to conduct pilot studies to promote the service to those who have never used it even if it could be for free. This could help service providers to raise awareness to these farmers and speed up the AI adoption.

The results also show that the milk marketing system plays a big role in driving farmers' decision. Where access to formal markets drives farmers' decision to supplement their animals, keeping of exotic animals, adopting best husbandry practices and attaining high productivity. This is useful information for decision-makers to consider various options of helping farmers to have access to markets. This could be carried out through having service providers and inputs

suppliers diversifying their business by linking farmers to reliable markets. This may also assure the sustainability of their business.

Lastly, this study contributes to the livestock sector on breeding and feeding practices and general animal husbandry practices for small scale dairy farmers. Compared to previous studies, the current study involved a large number of farmers and covered a wide area of Sub-Saharan Africa; Ethiopia, Kenya, Tanzania, and Uganda.

The analysis of multiple decisions helped to draw a big picture of the dairy industry of small-scale farmers. The dairy industry is also considered to be one of the complex sectors. As a result, there is a need for robust decision-support tools that would help to comprehensively identify and understand the big risks and opportunities, while examining multiple strategic options under different scenarios to pick the best solution. The current study assisted to show how different factors interact by transforming these into dynamic simulation models that analyze many more strategic options even more rapidly.

5.2 Recommendations

The results of this study, however, are subject to certain limitations. First, our sample is restricted to small scale dairy farmers of four countries in Sub Saharan Africa including Ethiopia, Kenya, Tanzania, and Uganda. Therefore, further work comparing the results on extended datasets from other countries would be beneficial. More studies need to be conducted for large and medium-size farmers in other countries and regions. The second potential limitation relates to the fact that the developed models were based on the data collected in 2015-2016 where the recommendation given in this study may not reflect current practices. Hence, in the future, the models need to be updated because the outcomes for a farm are likely to change over time. For example, there can be new technology or services introduced to farmers or change in farm and institution settings.

Also, the data that was used in this study was based on cross sectional survey which limits decision-makers to monitor an outcome. Thus, it can be difficult to analyze the behavior of farmers over a period of time or in determine cause and effect. To address this issue, one may set a system that can continue collecting data to create viable tests. For example, conducting of informative analysis of those farmers that have switched from one practice/adoption to another

and look at which features were predictive of the switch. In this way, policymakers can understand what compels farmers to change and design policies accordingly.

Lastly, despite this limitation, the achieved outcomes remain significant to the studied area including Ethiopia, Kenya, Tanzania and Uganda and during implementation, all rules defined must be observed such as the use of the appropriate value of K, maximum value of a number of animals or animal productivity (milk production) to be predicted.

REFERENCE

Ali, I., Cawkwell, F., Green, S., & Dwyer, N. (2014). Application of statistical and machine learning models for grassland yield estimation based on a hypertemporal satellite remote sensing time series. In *2014 IEEE Geoscience and Remote Sensing Symposium* (pp. 5060–5063). IEEE. https://doi.org/10.1109/IGARSS.2014.6947634.

Alsaaod, M., Römer, C., Kleinmanns, J., Hendriksen, K., RoseMeierhöfer, S., Plümer, L., & Büscher, W. (2012). Electronic detection of lameness in dairy cows through measuring pedometric activity and lying behavior. *Applied Animal Behaviour Science*, *142*(3–4), 134–141. https://doi.org/10.1016/J.Applanim.2012.10.001.

Amrine, D., White, B., & Larson, R. (2014). Comparison of classification algorithms to predict outcomes of feedlot cattle identified and treated for bovine respiratory disease. *Computers and Electronics in Agriculture*, *105*, (9–19), https://doi.org/10.1016/J. Compag. 2014.04.009.

AU-IBAR. (2015). *Livestock identification and Recording in Africa: Challenges, opportunities and options*.

Awasthi, A., Awasthi, A., Riordan, D., Walsh, J., Awasthi, A., Awasthi, A., … Walsh, J. (2016). Non-Invasive Sensor Technology for the Development of a Dairy Cattle Health Monitoring System. *Computers*, *5*(4), 23. https://doi.org/10.3390/computers 5040023.

Baltenweck, I., Ouma, R., Anunda, F., Mwai, O., & Romney, D. (2004). Artificial or natural insemination: The demand for breeding services by smallholders. In *ninth KARI (Kenya Agricultural Research Institute) annual scientific conference and agricultural research forum, Nairobi, Kenya*.

Barker, Z., Vázquez Diosdado, J., Codling, E., Bell, N., Hodges, H., Croft, D., & Amory, J. (2018). Use of novel sensors combining local positioning and acceleration to measure feeding behavior differences associated with lameness in dairy cattle. *Journal of Dairy Science*, *101*(7), 6310–6321. https://doi.org/10.3168/JDS.2016-12172.

Bayerni, P. (2012). Science and Technology for Livestock Value Chain Development: A Focus on Artificial Insemination. In *Knowledge for development:knowledge.cta*.

Benaissa, S., Tuyttens, F., Plets, D., Pessemier, T., Trogh, J., Tanghe, E., … Sonck, B. (2017). On the use of on-cow accelerometers for the classification of behaviours in dairy barns.*Research in Veterinary Science*. https://doi.org/10.1016/J.RVSC.2017.10.005

Borchers, M., Chang, Y., Proudfoot, K., Wadsworth, B., Stone, A., & Bewley, J. (2017). Machine-learning-based calving prediction from activity, lying, and ruminating behaviors in dairy cattle. *Journal of Dairy Science*, *100*(7), 5664–5674. https://doi.org/10.3168/JDS.2016-11526.

Borowska, A., Szwaczkowski, T., Kamiński, S., Hering, D., Kordan, W., & Lecewicz, M. (2018). Identification of genome regions determining semen quality in Holstein-Friesian bulls using information theory. *Animal Reproduction Science*, *192*, 206–215. https://doi.org/10.1016/J.Anireprosci.2018.03.012.

Brester, G., Marsh, J., & Plain, R. (2003). International red meat trade. *The Veterinary Clinics of North America. Food Animal Practice*, *19*(2), 493–518.

Brooks-pollock, E., Jong, M., Keeling, M., Klinkenberg, D., & Wood, J. (2015). Eight challenges in modelling infectious livestock diseases. *Epidemics*, *10*, 1–5. https://doi.org/10.1016/j.epidem.2014.08.005.

Caporale, V., Giovannini, A., Francesco, C., & Calistri, P. (2001). Importance of the traceability of animals and animal products in epidemiology. *Rev. Sci. Tech. Off. Int. Epiz*, *20*(2), 372–378.

Caraviello, D., Weigel, K., Craven, M., Gianola, D., Cook, N., Nordlund, K., … Wiltbank, M. C. (2006). Analysis of Reproductive Performance of Lactating Cows on Large Dairy Farms Using Machine Learning Algorithms. *Journal of Dairy Science*, *89*(12), 4703–4722. https://doi.org/10.3168/JDS.S0022-0302(06)72521-8.

Caroline, F., Luke, O., Michael, D., John, D., Laurence, S., & Stephen, B. (2017). The creation and evaluation of a model predicting the probability of conception in seasonal-calving, pasture-based dairy cows. *Journal of Dairy Science*, *100*(7), 5550–5563. https://doi.org/10.3168/JDS.2016-11830.

Chagunda, M., Msiska, A., Wollny, C., Tchale, H., & Banda, J. (2006). An analysis of smallholder farmers' willingness to adopt dairy performance recording in Malawi. *Livestock Research for Rural Development*, *18*(5).

Chebo, C., & Alemayehu, K. (2012). Trends of cattle genetic improvement programs in Ethiopia: Challenges and opportunities. *Livestock Research for Rural Development*, *24*(7).

Chelotti, J., Vanrell, S., Galli, J., Giovanini, L., & Rufiner, H. (2018). A pattern recognition approach for detecting and classifying jaw movements in grazing cattle. *Computers and Electronics in Agriculture*, *145*, 83–91. https://doi.org/10.1016/J. Compag.2017.12.013.

Chickering, D., & Heckerman, D. (2000). *A Decision Theoretic Approach to Targeted Advertising*. *Uncertainty in Artificial Intelligence proceedings*.

Chupin, D., & Thibier, M. (1995). Survey of the present status of the use of artificial insemination in developed countries. *World Animal Review*, *82*, 58–68.

Cook, J., & Green, M. (2016). Use of early lactation milk recording data to predict the calving to conception interval in dairy herds. *Journal of Dairy Science*, *99*(6), 4699–4706. https://doi.org/10.3168/JDS.2015-10264.

Dadzie, K., Amponsah, D., Dadzie, C., & Winston, E. (2017). How Firms Implement Marketing Strategies in Emerging Markets: An Empirical Assessment of The 4A Marketing Mix Framework. *Journal of Marketing Theory and Practice*, *25*(3), 234–256. https://doi.org/10.1080/10696679.2017.1311220.

Degenhardt, F., Seifert, S., & Szymczak, S. (2017). Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbx124.

Djedouboum, A., Abba Ari, A., Gueroui, A., Mohamadou, A., & Aliouat, Z. (2018). Big Data Collection in Large-Scale Wireless Sensor Networks. *Sensors (Basel, Switzerland)*, *18*(12). https://doi.org/10.3390/s18124474.

Dolecheck, K., Silvia, W., Heersche, G., Chang, Y., Ray, D., Stone, A., … Bewley, J. (2015). Behavioral and physiological changes around estrus events identified using multiple automated monitoring technologies. *Journal of Dairy Science*, *98*(12), 8723–8731. https://doi.org/10.3168/JDS.2015-9645.

Dórea, J., Rosa, G., Weld, K., & Armentano, L. (2018). Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows. *Journal of Dairy Science*, *101*(7), 5878–5889. https://doi.org/10.3168/JDS.2017-13997.

Dudafa, U. (2013). Record Keeping Among Small Farmers in Nigeria: Problems and Prospects. *International Journal of Scientific Research in Education*, *6*(2), 214–220.

Dutta, R., Smith, D., Rawnsley, R., Bishop-Hurley, G., Hills, J., Timms, G., & Henry, D. (2015). Dynamic cattle behavioural classification using supervised ensemble classifiers. *Computers and Electronics in Agriculture*, *111*, 18–28. https://doi.org/10.1016/J. compag.2014.12.002.

Ebrahimie, E., Ebrahimi, F., Ebrahimi, M., Tomlinson, S., & Petrovski, K. (2018). Hierarchical pattern recognition in milking parameters predicts mastitis prevalence. *Computers and Electronics in Agriculture*, *147*, 6–11. https://doi.org/10.1016/J. compag.2018.02.003.

Eric Conn. (2018). The Power of Cellular Technologies in IoT | IoT For All.

FAO. (2016). *The Global Dairy Sector: Facts*.

Fawcett, T. (2015). Mining the Quantified Self: Personal Knowledge Discovery as a Challenge for Data Science. *Big Data*, *3*(4), 249–266. https://doi.org/10.1089/big.2015.0049.

Fenlon, C., O'Grady, L., Mee, J., Butler, S., Doherty, M., & Dunnion, J. (2017). A comparison of 4 predictive models of calving assistance and difficulty in dairy heifers and cows. *Journal of Dairy Science*, *100*(12), 9746–9758. https://doi.org/10.3168/JDS.2017-12931.

Foote, R. (1996). Dairy cattle reproductive physiology research and management—Past progress and future prospects. *Journal of Dairy Science, Elsevier*, *79*, 980–990.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*.

Genuer, R., Poggi, J., & TuleauMalot, C. (2010). Variable selection using Random Forests. *Pattern Recognition Letters*.

Gimeno, E. (2003). The organisation and future development of Veterinary Services in Latin America. *Revue Scientifique et Technique (International Office of Epizootics)*, *22*(2), 449–461.

Goyache, F., Díez, J., López, S., Pajares, G., Santos, B., Fernández, I., & Prieto, M. (2005). Machine Learning as an aid to management decisions on high somatic cell counts in dairy farms. *Archives Animal Breeding*, *48*(2), 138–148. https://doi.org/10.5194/aab-48-138-2005.

Grace, D. (2013). The impact of animal disease on human hunger and health.

Grossberg, S. (2017). Towards solving the hard problem of consciousness: The varieties of brain resonances and the conscious experiences that they support. *Neural Networks*, *87*, 38–95. https://doi.org/10.1016/J.Neunet.2016.11.003.

Hadad, H., Mahmoud, H., & Mousa, F. (2015). Bovines Muzzle Classification Based on Machine Learning Techniques. *Procedia Computer Science*, *65*, 864–871. https://doi.org/10.1016/J.Procs.2015.09.044.

Hansson, H., & Lagerkvist, C. (2016). Dairy farmers' use and non-use values in animal welfare: Determining the empirical content and structure with anchored best-worst scaling. *Journal of Dairy Science*, *99*(1), 579–592. https://doi.org/10.3168/JDS.2015-9755.

Heikkilä, A., Myyrä, S., & Pietola, K. (2012). *Effects of Economic Factors on Adoption of Robotics and Consequences of Automation for Productivity Growth of Dairy Farms*.

Hempstalk, K., McParland, S., & Berry, D. (2015a). Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows. *Journal of Dairy Science*, *98*(8), 5262–5273. https://doi.org/10.3168/JDS.2014-8984.

Hempstalk, K., McParland, S., & Berry, D. (2015b). Machine learning algorithms for the prediction of conception success to a given insemination in lactating dairy cows. *Journal of Dairy Science*, *98*(8), 5262–5273. https://doi.org/10.3168/jds.2014-8984.

Hermans, K., Waegeman, W., Opsomer, G., Van Ranst, B., De Koster, J., Van Eetvelde, M., & Hostens, M. (2017). Novel approaches to assess the quality of fertility data stored in dairy herd management software. *Journal of Dairy Science*, *100*(5), 4078–4089. https://doi.org/10.3168/JDS.2016-11896.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. https://doi.org/10.1198/106186006X133933.

ILRI. (2018). Functions, objectives and instruments of policy.

Imandoust, S. B., & Bolandraftar, M. (2013). *Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background*. *Journal of Engineering Research and Applications www.ijera.com* (Vol. 3).

ISO. (1995). International Organization for Standardization (ISO).

Johnny, D. (2013). Record Keeping Among Small Farmers in Nigeria: Problems and Prospects. *International Journal of Scientific Research in Education*, *6*(62), 214–220.

Joseph Byrum. (2018). Agriculture's need for analytics and IoT - Informs.

Kabunga, N. (2014). Adoption and Impact of Improved Cow Breeds on Household Welfare and Child Nutrition Outcomes: Empirical Evidence from Uganda. *88th Annual Conference, April 9-11, 2014, AgroParisTech, Paris, France*.

Kamphuis, C., Mollenhorst, H., Feelders, A., Pietersma, D., & Hogeveen, H. (2010). Decision-tree induction to detect clinical mastitis with automatic milking. *Computers and Electronics in Agriculture*, *70*(1), 60–68. https://doi.org/10.1016/J.Compag. 2009.08.012.

Kanui, T., & Ikusya, K. (2016). *Innovativeness and Adaptations: The Way forward for Small scale Peri-Urban Dairy Farmers in Semi-Arid Regions of South Eastern Kenya* (Vol. 3).

Keegan, T., Cunningham, S., & Apperley, M. (1995). An investigation into the use of machine learning for determining oestrus in cows. *Hamilton, New Zealand: University of Waikato, Department of Computer Science.*

Khare, A., Jeon, M., Sethi, I., & Xu, B. (2017). Machine Learning Theory and Applications for Healthcare. *Journal of Healthcare Engineering*, *2017*, 5263570. https://doi.org/10.1155/2017/5263570.

Kotsiantis, S. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4.

Kumar, S., Pandey, A., Sai Ram Satwik, K., Kumar, S., Singh, S., Singh, A. , & Mohan, A. (2018). Deep learning framework for recognition of cattle using muzzle point image pattern. *Measurement*, *116*, 1–17. https://doi.org/10.1016/J.Measurement.2017.10.064.

Lemma, H., Mengistu, A., Kuma, T., & Kuma, B. (2018). Improving milk safety at farm-level in an intensive dairy production system: relevance to smallholder dairy producers. *Food Quality and Safety*, *2*(3), 135–143. https://doi.org/10.1093/fqsafe/fyy009.

Lior, R. (2014). *Data mining with decision trees: theory and applications*.

Loyola, D., Pedergnana, M., & Gimeno, S. (2016). Smart sampling and incremental function learning for very large high dimensional data. *Neural Networks*, *78*, 75–87. https://doi.org/10.1016/J.NEUNET.2015.09.001.

Ma, W., Fan, J., Li, Q., & Tang, Y. (2018). A raw milk service platform using BP Neural Network and Fuzzy Inference. *Information Processing in Agriculture*. https://doi.org/10.1016/J.INPA.2018.04.001.

Mammadova, N., & Keskin, I. (2013). Application of the support vector machine to predict subclinical mastitis in dairy cattle. *The Scientific World Journal*, *2013*, 603897. https://doi.org/10.1155/2013/603897.

Mbwambo, N., Nigussie, K., & Stapleton O, J. I. (2017). *Dairy development in the Tanzanian Livestock master plan*.

McKean, J. (2001). The importance of traceability for public health and consumer protection. *International Office of Epizootics*, *20*(2), 363–371].

Mitchell, R., Sherlock, R., & Smith, L. (1996). An investigation into the use of machine learning for determining oestrus in cows. *Computers and Electronics in Agriculture*, *15*(3), 195–213. https://doi.org/10.1016/0168-1699(96)00016-6.

MoF, M. (2016). *Agriculture Sector Strategic Plan (Uganda)*.

Mugisha, A., Kayiizi, V., Owiny, D., & Mburu, J. (2014). Breeding services and the factors influencing their use on smallholder dairy farms in central Uganda. *Veterinary Medicine International*, *2014*, 169380. https://doi.org/10.1155/2014/169380.

Murage, A., & Ilatsia, E. (2011). Factors that determine use of breeding services by smallholder dairy farmers in Central Kenya. *Tropical Animal Health and Production*, *43*(1), 43:199–207. https://doi.org/10.1007/s11250-010-9674-3.

Mutarutwa, N. (2014). The impact of the Girinka one cow per poor family program on household income in Gatsibo District, Rwanda.

Mwanga, G., Mujibi, F., Yonah, Z., & Chagunda, M. (2018). Multi-country investigation of factors influencing breeding decisions by smallholder dairy farmers in sub-Saharan Africa. *Tropical Animal Health and Production*, *51*(2), 1–15. https://doi.org/10.1007/s11250-018-1703-7.

Nancy Morgan. (2018). Dairy policies and sector planning.

Nasirahmadi, A., Edwards, S. A., & Sturm, B. (2017). Implementation of machine vision for detecting behaviour of cattle and pigs. *Livestock Science*, *202*, 25–38. https://doi.org/10.1016/J.LIVSCI.2017.05.014.

Nell, J., Schiere, H., & Bol, S. (2014). *Quick scan Dairy Sector Tanzania*.

Parker Gaddis, K., Cole, J., Clay, J., & Maltecca, C. (2016). Benchmarking dairy herd health status using routinely recorded herd summary data. *Journal of Dairy Science*, *99*(2), 1298–1314. https://doi.org/10.3168/JDS.2015-9840.

Pica-Ciamarra, U., Otte, J., & Martini, C. (2010). *A Living from Livestock Pro-Poor Livestock Policy Initiative Livestock Sector Policies and Programmes in Developing Countries A Menu for Practitioners*.

Pietersma, D., Lacroix, R., Lsfebvre, D., & Wade, K. (2003). Induction and evaluation of decision trees for lactation curve analysis. *Computers and Electronics in Agriculture*, *38*(1), 19–32. https://doi.org/10.1016/S0168-1699(02)00105-9.

Rahman, A., Smith, D., Little, B., Ingham, A., Greenwood, P., & Bishop-Hurley, G. (2018). Cattle behaviour classification from collar, halter, and ear tag sensors. *Information Processing in Agriculture*, *5*(1), 124–133. https://doi.org/10.1016/ J.INPA.2017.10.001.

Richards, S., VanLeeuwen, J., Shepelo, G., Gitau, G., Kamunde, C., Uehlinger, F., & Wichtel, J. (2015). Associations of farm management practices with annual milk sales on smallholder dairy farms in Kenya. *Veterinary World*, *8*(1), 88–96. https://doi.org/10.14202/vetworld.2015.88-96.

Ristoski, P., & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on the World Wide Web*, *36*, 1–22. https://doi.org/10.1016/J.Websem.2016.01.001.

Rodrigues, F., & Ferreira, B. (2016). Product Recommendation based on Shared Customer's Behaviour. *Procedia Computer Science*, *100*, 136–146. https://doi.org/10.1016/ J.PROCS.2016.09.133.

Roelofs, J., López-Gatius, F., Hunter, R., van Eerdenburg, F., & Hanzen, C. (2010). When is a cow in estrus? Clinical and practical aspects. *Theriogenology*, *74*(3), 327–344. https://doi.org/10.1016/j.theriogenology.2010.02.016.

Roland, L., Lidauer, L., Sattlecker, G., Kickinger, F., Auer, W., Sturm, V., … Iwersen, M. (2018). Monitoring drinking behavior in bucket-fed dairy calves using an ear-attached tri-axial accelerometer: A pilot study. *Computers and Electronics in Agriculture*, *145*, 298–301. https://doi.org/10.1016/J.Compag.2018.01.008.

Rutten, C., Steeneveld, W., Vernooij, J., Huijps, K., Nielen, M., & Hogeveen, H. (2016). A prognostic model to predict the success of artificial insemination in dairy cows based on readily available data. *Journal of Dairy Science*, *99*(8), 6764–6779. https://doi.org/10.3168/JDS.2016-10935.

Saint-Dizier, M., & Chastant-Maillard, S. (2018). Potential of connected devices to optimize cattle reproduction. *Theriogenology*, *112*, 53–62. https://doi.org/10.1016/ J.Theriogenology.2017.09.033.

Saleh, S., David, P., Jerry, G., Victor, C., Paul, F., & Kent, W. (2014). Prediction of insemination outcomes in Holstein dairy cattle using alternative machine learning algorithms. *Journal of Dairy Science*, *97*(2), 731–742. https://doi.org/10.3168/JDS.2013-6693.

Santoni, M., Sensuse, D., Arymurthy, A., & Fanany, M. (2015). Cattle Race Classification Using Gray Level Co-occurrence Matrix Convolutional Neural Networks. *Procedia Computer Science*, *59*, 493–502. https://doi.org/10.1016/J.PROCS.2015.07.525

SAS, S. (2003). Version. "9.4 [Computer Program]." SAS Institute, Cary, NC.

Satterthwaite, D., McGranahan, G., & Tacoli, C. (2010). Urbanization and its implications for food and farming. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *365*(1554), 2809–2820. https://doi.org/10.1098/rstb.2010.0136.

Schefers, J., Weigel, K., Rawson, C., Zwald, N., & Cook, N. (2010). Management practices associated with conception rate and service rate of lactating Holstein cows in large, commercial dairy herds. *Journal of Dairy Science*, *93*(4), 1459–1467. https://doi.org/10.3168/JDS.2009-2015.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

Scrucca, L., Fop, M., Murphy, B., & Raftery, A. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, *8*(1), 289–317.

SEMTECH. (2017). *LoRa Application Brief Smart Agriculture Semtech's LoRa Technology Enables Smart Agriculture*.

Shahinfar, S., Kalantari, A., Cabrera, V., & Weigel, K. (2014). Short communication: Prediction of retention pay-off using a machine learning algorithm. *Journal of Dairy Science*, *97*(5), 2949–2952. https://doi.org/10.3168/JDS.2013-7373.

Shahinfar, S., Mehrabani-Yeganeh, H., Lucas, C., Kalhor, A., Kazemian, M., & Weigel, K. (2012). Prediction of Breeding Values for Dairy Cattle Using Artificial Neural Networks and Neuro-Fuzzy Systems. *Computational and Mathematical Methods in Medicine*, *2012*, 1–9. https://doi.org/10.1155/2012/127130.

Shahriar, M., Smith, D., Rahman, A., Freeman, M., Hills, J., Rawnsley, R., … Bishop-Hurley, G. (2016). Detecting heat events in dairy cows using accelerometers and unsupervised learning. *Computers and Electronics in Agriculture*, *128*, 20–26. https://doi.org/10.1016/J.Compag.2016.08.009.

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., & Khovanova, N. (2017). Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. *Biomedical Signal Processing and Control*. https://doi.org/10.1016/J.BSPC.2017.01.012.

Shine, P., Murphy, M., Upton, J., & Scully, T. (2018). Machine-learning algorithms for predicting on-farm direct water and electricity consumption on pasture based dairy farms. *Computers and Electronics in Agriculture*, *150*, 74–87. https://doi.org/10.1016/J.Compag.2018.03.023.

Singh, A., Halgamuge, M., & Lakshmiganthan, R. (2017). Impact of Different Data Types on Classifier Performance of Random Forest, Naïve Bayes, and K-Nearest Neighbors Algorithms. *IJACSA) International Journal of Advanced Computer Science and Applications*, *8*(12).

Smith, D., Rahman, A., Bishop-Hurley, G., Hills, J., Shahriar, S., Henry, D., & Rawnsley, R. (2016). Behavior classification of cows fitted with motion collars: Decomposing multi-class classification into a set of binary problems. *Computers and Electronics in Agriculture*, *131*, 40–50. https://doi.org/10.1016/J.Compag.2016.10.006.

Steinfeld, H., & Mack, S. (2019). Livestock development strategies. Retrieved November 4, 2019, from http://www.fao.org/3/V8180T/v8180T0a.htm.

Tefera, S., Lagat, J., & Bett, H. (2014). Determinants of Artificial Insemination Use by Smallholder Dairy Farmers in Lemu-Bilbilo District, Ethiopia. *International Journal of African and Asian Studies*, *4*.

Tharwat, A., Gaber, T., & Hassanien, A. (2014). Cattle Identification Based on Muzzle Images Using Gabor Features and SVM Classifier (pp. 236–247). Springer, Cham. https://doi.org/10.1007/978-3-319-13461-1_23.

The World Bank, F. (2014). *Investing in the livestock sector: Why Good Numbers Matter*.

Thiermann, A. (2005). Globalization, international trade and animal health: the new roles of OIE. *Preventive Veterinary Medicine*, *67*(2–3), 101–108.

Thornton, P. (2010). Livestock production: recent trends, future prospects. *The Royal Society*, *365*(1554), 2853–2867. https://doi.org/10.1098/rstb.2010.0134.

Tongeren, F. (2008). Agricultural Policy Design and Implementation A Synthesis. *OECD Food, Agriculture and Fisheries*, *7*. https://doi.org/10.1787/243786286663.

Ugo Pica-Ciamarra, Derek Baker, Nancy Morgan, Alberto Zezza, Carlo Azzarri, Cheikh Ly, … Joseph Sserugga. (2014). *Investing in the livestock sector Why Good Numbers Matter A Sourcebook for Decision Makers on How to Improve Livestock Data*.

UNPD, (United Nations Population Division). (2008). *The 2006 revision and world urbanization prospects: the 2005 Revision. Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat, World Population Prospects*.

VanLeeuwen, J., Mellish, T., Walton, C., Kaniaru, A., Gitau, R., Mellish, K., … Wichtel, J. (2012). Management, productivity and livelihood effects on Kenyan smallholder dairy farms from interventions addressing animal health and nutrition and milk quality. *Tropical Animal Health and Production*, *44*(2), 231–238. https://doi.org/10.1007/s11250-011-0003-2.

Viazzi, S., Bahr, C., Schlageter-Tello, A., Van Hertem, T., Romanini, C., Pluk, A., … Berckmans, D. (2013). Analysis of individual classification of lameness using automatic measurement of back posture in dairy cattle. *Journal of Dairy Science*, *96*(1), 257–266. https://doi.org/10.3168/JDS.2012-5806.

Wei, Z., Wang, J., & Zhang, X. (2013). Monitoring of quality and storage time of unsealed pasteurized milk by voltammetric electronic tongue. *Electrochimica Acta*, *88*, 231–239. https://doi.org/10.1016/J.Electacta.2012.10.042.

Williams, M., Mac Parthaláin, N., Brewer, P., James, W., & Rose, M. (2016). A novel behavioral model of the pasture-based dairy cow from GPS data using data mining and machine learning techniques. *Journal of Dairy Science*, *99*(3), 2063–2075. https://doi.org/10.3168/JDS.2015-10254.

Yazdanbakhsh, O., Zhou, Y., & Dick, S. (2017). An intelligent system for livestock disease surveillance. *Information Sciences*, *378*, 26–47. https://doi.org/10.1016/ J.INS.2016.10.026.

Yeomans, M. (2015). What Every Manager Should Know About Machine Learning. Retrieved from https://hbr.org/2015/07/what-every-manager-should-know-about-machine-learning

Zaborski, D., Proskura, W., Grzesiak, W., Szatkowska, I., & Jędrzejczak-Silicka, M. (2017). Use of random forest for dystocia detection in dairy cattle. *Appl Agric Forestry Res*, *1484*(367), 2017147–2017154. https://doi.org/10.3220/LBF1515508151000.

Zepeda, C., Salman, M., Thiermann, A., Kellar, J., Rojas, H., & Willeberg, P. (2005). The role of veterinary epidemiology and veterinary services in complying with the World Trade Organization SPS agreement. *Preventive Veterinary Medicine*, *67*(2–3), 125–140.

Zhang, L., Zhang, X., Ni, L., Xue, Z., Gu, X., & Huang, S. (2014). Rapid identification of adulterated cow milk by non-linear pattern recognition methods based on near infrared spectroscopy. *Food Chemistry*, *145*, 342–348. https://doi.org/10.1016/ J.Foodchem. 2013.08.064.

Zhao, K., Bewley, J., He, D., & Jin, X. (2018). Automatic lameness detection in dairy cattle based on leg swing analysis with an image processing technique. *Computers and Electronics in Agriculture*, *148*, 226–236. https://doi.org/10.1016/J. Compag.2018.03.014.

Appendix 1: Features that were tested for model development

This appendix lists the variables that were tested for model development, categorized in different categories including Demographic Information, Farm Characteristics, Institution settings, Cost Incurred for managing a farm and animals, breeding technologies, feeding systems, farm income, farm management practices and animal health.

Table 18: List of variables that were tested for model's development, extracted from collected data.

Demographic Information

1. Farmers education
2. Total number of children in the Household
3. Farmers experience in dairy
4. Study sites (Regions)
5. Farmer involvement to farmers groups

Farm Characteristics

1. Total land size
2. Heard size.
3. Source of water (tap, River, Borehole, Pan, Pond and Rain)
4. Number of milking cows
5. Lactation Length
6. Number of exotic animals
7. Number of local animals
8. Type of crops grown (cash crops, food crops, fodder crops and grazing grasses)
9. Number of bulls on the farm
10. Number of castrated adult male on the farm
11. Number of immature males on the farm
12. Number of heifers on the farm
13. Number of female calves on the farm
14. Number of male calves on the farm
15. Frequency of watering animals
16. Frequency of treating animals

17. Total number of labors
18. Labors working hours

Institution settings

1. Time to market
2. Distance Milk Buyer
3. Time taken to transport milk
4. Distance to market
5. Availability of vaccination service
6. Water Availability

Cost Incurred for managing a farm and animals

1. Average cost for breeding method
2. Cost for water
3. Transportation cost for milk

Breeding services

1. Breeding service provider (private, cooperation, government, individual and neighbor)
2. Distance to service provider
3. Preferred breeding method
4. Breeding method recently calved
5. Availability of preferred breeding service
6. Number of repeats for the success conception rate

Feeding systems

1. Source of purchased fodder (supplier, local trader, dairy cooperative and fodder inputs)
2. Number of months purchased fodder
3. Source of crop residue (own farms, Scavenged people's farms, purchased and processing machine)
4. Source of purchased crop residue (neighbor, supplier, local trader, dairy cooperative and input supplier)
5. Number of months used crop residue
6. Concentrate usage

7. Preferred buyers (Individual consumers, Private milk-traders, Dairy co-op/ group with, chilling plants, Milk collection center, Hotels and other institutions)
8. Feeding system used on the farm during the rainy season (Only grazing, mainly grazing with some stall feeding, Mainly, stall feeding with some grazing, Only stall feeding (zero grazing), Transhumance some animals, Transhumance all animals)
9. Feeding system used on the farm during dry season (Only grazing, mainly grazing with some stall feeding, Mainly, stall feeding with some grazing, Only stall feeding (zero grazing), Transhumance some animals, Transhumance, all animals)

Farm income

1. Liters Sold
2. Total income from crops
3. Processed milk products (fermented milk, butter, ghee, ice cream, cheese, yoghurt and Reserved for Consumption)

Farm management practices

1. Keeping of records
2. Types of records kept by a farmer (breeding, calving, feeding, health, milk production, growth, sales and traceability)
3. Record usage (identity, extension officer, self-evaluation, sales and management)
4. Animal identification system used (tags, name, markers, notching, tattooing, none)

Animal health

1. Times vaccinated
2. Deworming times

Appendix 2: Sample R codes that were employed for features and model's selection

This appendix illustrates the sample R codes that were employed for features engineering and model's evaluation. The sample code was used in modeling farmers decisions regard to the use of concentrate. All models developed in this study followed the same approach.

Table 19: Sample R code that was used for features and model selection.

```
library(mclust)
library(Boruta)
library(boot)
library(caret)
library(rpart)
library(randomForest)

#setwd("C:/Users/mwanga/GoogleDrive/Data
 analysis/R_modeling/Repetition/Repetition4/Models_all_countries/Feeding")

country.names<- c("Kenya","Tanzania","Ethiopia","Uganda")

all.LM<- list()
all.boruta<- list()
all.RF<- list()

# k-fold crossvalidation
ctrl<- trainControl(method = "cv", number = 10, savePredictions = TRUE)

# ====================COMPUTE FEATURES====================
all.accuracies<- c() # accuracies for all countries
for(iin1:length(country.names)){
accuracies<- c() # accuracies for this country

  print(paste("Working on: ",country.names[i],"_modeling_concetrate.csv", sep =""))
# Read the data
country<- read.csv(paste(country.names[i], "_modeling_concetrate.csv", sep = ""))
#country<-round(country, 1)

#country<-na.omit(country)
# Class must be as a factor
# country[, ncol(country)] <- as.factor(country[, ncol(country)])
# 3-fold crossvalidation
ctrl<- trainControl(method = "cv", number = 10, savePredictions = TRUE)

# # Copmute Boruta features
boruta_output<- Boruta(country[,ncol(country)] ~ ., data=country[,-ncol(country)],
 doTrace=2)
```

```
start<-attStats(boruta_output)
Impfeatures<- start[start$decision == "Confirmed",c('meanImp','decision')]
BRtopVar<- head(Impfeatures[order(-Impfeatures$meanImp),],n=15)
boruta.features<-which(BRtopVar$decision== "Confirmed")

# # Copmute Boruta features
boruta_output<- Boruta(country[,ncol(country)] ~ ., data=country[,-ncol(country)],
 doTrace=2)
start<-attStats(boruta_output)

Impfeatures<- start[c('meanImp','decision')]

BRtopVar<- Impfeatures[order(-Impfeatures$meanImp),]
boruta.features<-head(which(BRtopVar== "Confirmed"),15)
  print(boruta.features)

boruta.features<-names(head(which(Impfeatures$decision== "Confirmed"),15))

Impfeatures<- start[start$decision == "Confirmed",c('meanImp','decision')]
BRtopVar<- head(Impfeatures[order(-Impfeatures$meanImp),],n=15)
boruta.features<-names( which(Impfeatures$decision== "Confirmed"))


# # Compute Random Forest features
RF_output<- randomForest(buy_concentrate ~ ., data=country,importance=T,trControl =
 ctrl)
varImportance<- as.data.frame(importance(RF_output))
RFtopVar<- head( varImportance[order(- varImportance$`%IncMSE`),],n=15)
RF.features<- which(RFtopVar$`%IncMSE`>1)


############################Logistic Models ############################
# Logistic on all vars
model.fit<- train(buy_concentrate ~., data=country, method="glm", family="binomial",
trControl = ctrl)
# print(summary(model.fit))
# Computing prediction accuracy
xx<- model.fit$pred
curr.acc<-mean(as.numeric(xx$pred>0.5)== xx[, 2])
accuracies<- c(accuracies, curr.acc)


# Logistic on signif vars
  summary(model.fit)
modcoef<- summary(model.fit)[["coefficients"]]
orderedVar<-as.data.frame(modcoef[order(modcoef[ , 4]), ])
LMtopVar<- head( orderedVar,n=15)
LM.signif<- which(LMtopVar$`Pr(>|z|)`<= 0.05)

data.new<- country[, LM.signif]
```

```
data.new$buy_concentrate<- country$buy_concentrate
model.fit<- train(buy_concentrate ~., data=data.new, method="glm", family="binomial",
trControl = ctrl)
# Computing prediction accuracy
xx<- model.fit$pred
curr.acc<-mean(as.numeric(xx$pred>0.5)== xx[, 2])
accuracies<- c(accuracies, curr.acc)

# Logistic on Boruta  features
data.new<- country[, boruta.features]
data.new$buy_concentrate<- country$buy_concentrate
model.fit<- train(buy_concentrate ~., data=data.new, method="glm", family="binomial",
trControl = ctrl)
# Computing prediction accuracy
xx<- model.fit$pred
curr.acc<-mean(as.numeric(xx$pred>0.5)== xx[, 2])
accuracies<- c(accuracies, curr.acc)


# Logistic on RF  features
data.new<- country[, RF.features]
data.new$buy_concentrate<- country$buy_concentrate
model.fit<- train(buy_concentrate ~., data=data.new, method="glm", family="binomial",
trControl = ctrl)
# Computing prediction accuracy
xx<- model.fit$pred
curr.acc<-mean(as.numeric(xx$pred>0.5)== xx[, 2])
accuracies<- c(accuracies, curr.acc)


all.accuracies<- rbind(all.accuracies, accuracies)
all.LM[[country.names[i]]] <- names(country)[LM.signif]
all.boruta[[country.names[i]]] <- names(country)[boruta.features]
all.RF[[country.names[i]]] <- names(country)[RF.features]
}

##Print features
sink("boruta_features.txt")
print(all.boruta)
sink()

sink("logistic_signif_features.txt")
print(all.LM)
sink()

sink("logistic_signif_features.txt")
print(all.RF)
sink()
row.names(all.accuracies) <- country.names
colnames(all.accuracies) <- c("All", "LM.signif", "Boruta","RF")
```

```
write.csv(all.accuracies, "accuracies_glm.csv")


##############################Decision Tree ############################

all.accuracies<- c() # accuracies for all countries
for(iin1:length(country.names)){
accuracies<- c() # accuracies for this country
  print(paste("Working on: ",country.names[i],"_modeling_concetrate.csv", sep =""))
# Read the data
country<- read.csv(paste(country.names[i], "_modeling_concetrate.csv", sep = ""))
country<-round(country, 1)

# 3-fold crossvalidation
ctrl<- trainControl(method = "cv", number = 10, savePredictions = TRUE)

###Decision tree on all variables
fitdecitiontree<- train(buy_concentrate ~ ., data=country, method="rpart",trControl = ctrl)
xx<- fitdecitiontree$pred
curr.acc<-mean(as.numeric(xx$pred>0.5)== xx[, 2])
accuracies<- c(accuracies, curr.acc)

###Decision tree for Logistic variables
ii<- which(names(country) %in% all.LM[[i]])
X<- country[, ii]
X$buy_concentrate<- country$buy_concentrate

fitdecitiontree<- train(buy_concentrate ~ ., data=X , method="rpart",trControl = ctrl)
xx<- fitdecitiontree$pred
curr.acc<-mean(as.numeric(xx$pred>0.5)== xx[, 2])
accuracies<- c(accuracies, curr.acc)

###Decision tree for Boruta variables
ii<- which(names(country) %in% all.boruta[[i]])
X<- country[, ii]
X$buy_concentrate<- country$buy_concentrate

fitdecitiontree<- train(buy_concentrate ~ ., data=X , method="rpart",trControl = ctrl)
xx<- fitdecitiontree$pred
curr.acc<-mean(as.numeric(xx$pred>0.5)== xx[, 2])
accuracies<- c(accuracies, curr.acc)

###Decision tree for RF variables
ii<- which(names(country) %in% all.RF[[i]])
X<- country[, ii]
X$buy_concentrate<- country$buy_concentrate

fitdecitiontree<- train(buy_concentrate ~ ., data=X , method="rpart",trControl = ctrl)
xx<- fitdecitiontree$pred
curr.acc<-mean(as.numeric(xx$pred>0.5)== xx[, 2])
```

```
accuracies<- c(accuracies, curr.acc)


# Computing prediction accuracy
all.accuracies<- rbind(all.accuracies, accuracies)
all.LM[[country.names[i]]] <- names(country)[LM.signif]
all.boruta[[country.names[i]]] <- names(country)[boruta.features]
all.RF[[country.names[i]]] <- names(country)[RF.features]
}

row.names(all.accuracies) <- country.names
colnames(all.accuracies) <- c("All", "LM.signif", "Boruta","RF")
write.csv(all.accuracies, "accuracies_DT.csv")


##################### Random forest #########################

all.accuracies<- c() # accuracies for all countries
for(iin1:length(country.names)){
accuracies<- c() # accuracies for this country
  print(paste("Working on: ",country.names[i],"_modeling_concetrate.csv", sep =""))
# Read the data
country<- read.csv(paste(country.names[i], "_modeling_concetrate.csv", sep = ""))
country<-round(country, 1)

###RF on all variables
fitRF<- randomForest(buy_concentrate ~ ., data=country,trControl = ctrl)
xx<- fitRF$predicted
xx<-as.numeric(xx>0.5)

curr.acc<- mean(country$buy_concentrate==xx)
accuracies<- c(accuracies, curr.acc)

###RF on Sign Variables
ii<- which(names(country) %in% all.LM[[i]])
XR<- country[, ii]
XR$buy_concentrate<- country$buy_concentrate

fitRF<- randomForest(buy_concentrate ~ ., data=XR ,trControl = ctrl)

xx<- fitRF$predicted
xx<-as.numeric(xx>0.5)
curr.acc<- mean(XR$buy_concentrate==xx)
accuracies<- c(accuracies, curr.acc)


###RF on Boruta
ii<- which(names(country) %in% all.boruta[[i]])
XR<- country[, ii]
XR$buy_concentrate<- country$buy_concentrate
```

```
fitRF<- randomForest(buy_concentrate ~ ., data=XR,trControl = ctrl)

xx<- fitRF$predicted
xx<-as.numeric(xx>0.5)
curr.acc<- mean(XR$buy_concentrate==xx)
accuracies<- c(accuracies, curr.acc)

###RF on RF Features
ii<- which(names(country) %in% all.RF[[i]])
XR<- country[, ii]
XR$buy_concentrate<- country$buy_concentrate

fitRF<- randomForest(buy_concentrate ~ ., data=XR,trControl = ctrl)

xx<- fitRF$predicted
xx<-as.numeric(xx>0.5)
curr.acc<- mean(XR$buy_concentrate==xx)
accuracies<- c(accuracies, curr.acc)


# Computing prediction accuracy
all.accuracies<- rbind(all.accuracies, accuracies)
all.LM[[country.names[i]]] <- names(country)[LM.signif]
all.boruta[[country.names[i]]] <- names(country)[boruta.features]
all.RF[[country.names[i]]] <- names(country)[RF.features]
}
row.names(all.accuracies) <- country.names
colnames(all.accuracies) <- c("All", "LM.signif", "Boruta","RF")
write.csv(all.accuracies, "accuracies_RF.csv")


# ###################### KNN ######################
# The following code is contingeoun in the previous code
all.accuracies<- c() # accuracies for all countries
all.ks<- c()
for(iin1:length(country.names)){
accuracies<- c() # accuracies for this country
current.ks<- c() # best k's for KNN
  print(paste("Working on: ",country.names[i],"_modeling_concetrate.csv", sep =""))
# Read the data
country<- read.csv(paste(country.names[i], "_modeling_concetrate.csv", sep = ""))
# Remove weights after reading them
country<-round(country, 1)

# Class must be as a factor
#country[, ncol(country)] <- as.factor(country[, ncol(country)])

ks<- as.data.frame(c(15:100))
  names(ks) <- "k"
```

```
# KNN on all features
knn.fit<- train(buy_concentrate ~., data=country, method = "knn", tuneGrid=ks,
trControl = ctrl, preProcess = c("center","scale"))
# Computing prediction accuracy
xx<- knn.fit$pred
xxx<- knn.fit$bestTune
curr.acc<- mean(xx[xx$k==xxx$k, 1] == xx[xx$k==xxx$k, 2])
accuracies<- c(accuracies, curr.acc)
current.ks<- c(current.ks, xxx$k)

# KNN on signif vars
ii<- which(names(country) %in% all.LM[[i]])
data.new<- country[, ii]
data.new$buy_concentrate<- country$buy_concentrate
knn.fit<- train(buy_concentrate ~., data=data.new, method = "knn", tuneGrid=ks,
trControl = ctrl, preProcess = c("center","scale"))
# Computing prediction accuracy
xx<- knn.fit$pred
xxx<- knn.fit$bestTune
curr.acc<- mean(xx[xx$k==xxx$k, 1] == xx[xx$k==xxx$k, 2])
accuracies<- c(accuracies, curr.acc)
current.ks<- c(current.ks, xxx$k)

# KNN on Boruta  features
ii<- which(names(country) %in% all.boruta[[i]])
data.new<- country[, ii]
data.new$buy_concentrate<- country$buy_concentrate
knn.fit<- train(buy_concentrate ~., data=data.new, method = "knn", tuneGrid=ks,
trControl = ctrl, preProcess = c("center","scale"))
# Computing prediction accuracy
xx<- knn.fit$pred
xxx<- knn.fit$bestTune
curr.acc<- mean(xx[xx$k==xxx$k, 1] == xx[xx$k==xxx$k, 2])
accuracies<- c(accuracies, curr.acc)
current.ks<- c(current.ks, xxx$k)

# KNN on RF  features
ii<- which(names(country) %in% all.RF[[i]])
data.new<- country[, ii]
data.new$buy_concentrate<- country$buy_concentrate
knn.fit<- train(buy_concentrate ~., data=data.new, method = "knn", tuneGrid=ks,
trControl = ctrl, preProcess = c("center","scale"))
# Computing prediction accuracy
xx<- knn.fit$pred
xxx<- knn.fit$bestTune
curr.acc<- mean(xx[xx$k==xxx$k, 1] == xx[xx$k==xxx$k, 2])
accuracies<- c(accuracies, curr.acc)
current.ks<- c(current.ks, xxx$k)
```

```
all.accuracies<- rbind(all.accuracies, accuracies)
all.ks<- rbind(all.ks, current.ks)
}

row.names(all.accuracies) <- country.names
colnames(all.accuracies) <- c("All", "LM.signif", "Boruta","RF")
write.csv(all.accuracies, "accuracies_knn.csv")

row.names(all.ks) <- country.names
colnames(all.ks) <- c("All", "LM.signif", "Boruta","RF")
write.csv(all.ks, "ks_knn.csv")
```

Appendix 3: Additional results

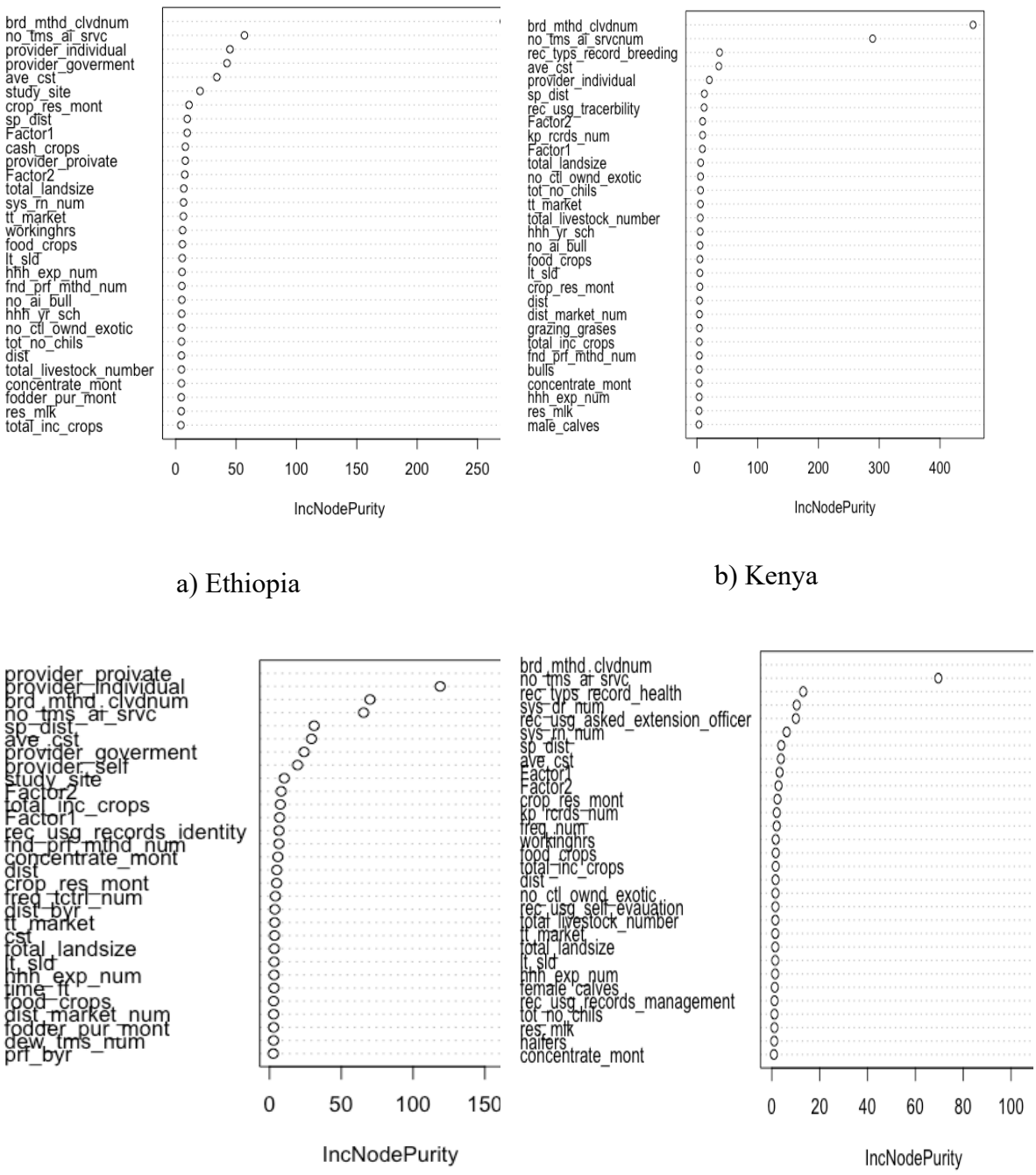
Table 20: Variables selected by linear models to be used in developing prediction model to predict farmers decision in regard to breeding method in Ethiopia, Kenya, Tanzania and Uganda

	Ethiopia			Kenya	
	Estimate	**Pr(>|t|)**		**Estimate**	**Pr(>|t|)**
brd_mthd_clvdnum	3.70E+00	5.63E-82	brd_mthd_clvdnum	4.43E+00	6.81E-109
study_sitebahir_dar	3.85E+00	7.48E-25	no_tms_ai_srvcnum	1.56E+00	4.23E-53
provider_individual	-2.33E+00	1.84E-07	provider_individual	-4.95E+00	2.70E-14
no_tms_ai_srvc	4.54E-01	9.64E-07	provider_cooperation	-7.88E-01	1.91E-05
study_siteasela_shed	2.38E+00	4.13E-06	fnd_prf_mthd_num	7.59E-01	3.11E-04
fodder_pur_mont	-1.56E-01	1.15E-05	provider_self	-2.55E+00	4.09E-04
provider_goverment	1.57E+00	3.13E-05	ave_cst	5.89E-04	1.02E-03
fnd_prf_mthd_num	7.99E-01	2.56E-04	ctl_id_tags	8.07E-01	1.51E-03
study_sitehawassa_shed	1.66E+00	3.71E-04	rec_typs_record_breeding	1.14E+00	2.72E-03
prf_byr4	1.20E+00	9.99E-04	rec_usg_records_identity	-1.13E+00	3.53E-03
status_num	-5.72E-01	1.67E-03	dist_market_num	1.04E-01	5.31E-03
provider_self	-1.67E+00	1.97E-03	rec_usg_records_sales	-1.00E+00	7.46E-03
freq_num	3.09E-01	9.97E-03	src_residue_own_farm	6.25E-01	1.04E-02
cash_crops	4.32E-01	1.04E-02	fodder_pur_mont	1.28E-01	2.23E-02
wh_local_trader	6.19E-01	1.43E-02	study_sitenorth_lift	-4.14E-01	3.30E-02
total_livestock_number	-9.00E-02	1.93E-02	provider_goverment	-7.36E-01	3.74E-02
no_ai_bull	2.31E-01	3.48E-02	src_residue_purchased	1.56E+00	3.84E-02
res_mlk	9.11E-02	4.22E-02	no_ai_bull	-2.39E-01	4.37E-02
spend	-7.09E-03	4.53E-02	rec_typs_record_milk	7.25E-01	5.90E-02
src_residue_purchased	5.55E-01	4.55E-02	cash_crops	-6.36E-02	6.96E-02
src_scavenged_peoples_farm	-1.60E+00	4.84E-02	tt_market	-7.12E-02	9.67E-02
concentrate_mont	5.75E-02	4.91E-02	haifers	5.52E-01	9.79E-02
rec_usg_self_evauation	-1.07E+00	6.18E-02	wh_feed_supplier	-1.08E+00	1.01E-01
rec_typs_record_culving	1.34E+00	6.47E-02	ctl_id_id_none	9.80E-01	1.18E-01
rec_usg_tracerbility	1.20E+00	6.67E-02	total_livestock_number	3.18E-02	1.21E-01
totalLlabour	-1.66E-01	6.92E-02	ctl_id_tattooing	-1.41E+00	1.24E-01
prf_byr1	6.04E-01	7.61E-02	src_water_rain	-3.97E-01	1.35E-01
src_water_rain	-3.21E+00	7.62E-02	wh_fodder_inputs	9.79E-01	1.40E-01
freq_tctrl_num	-5.12E-02	1.02E-01	crop_res_mont	-4.63E-02	1.44E-01
sys_dr_num3	-1.13E+00	1.10E-01	rec_usg_tracerbility	5.95E-01	1.49E-01
grazing_grases	-1.98E-01	1.12E-01	grazing_grases	-3.12E-02	1.59E-01
purchase_feed_suppplier	1.22E+00	1.15E-01	purchase_neighbor	-1.23E+00	1.60E-01
hhh_yr_sch	2.72E-02	1.18E-01	immature_males	4.69E-01	1.73E-01
ctl_id_id_none	-7.54E-01	1.19E-01	ctl_id_id_name	7.81E-01	1.84E-01
avail_num	3.37E-01	1.25E-01	no_ctl_ownd_exotic	-4.01E-01	2.13E-01

crop_res_mont	-3.71E-02	1.29E-01	src_water_tap	-3.90E-01	2.15E-01
kp_rcrds_num	-1.13E+00	1.33E-01	concentrate_mont	2.60E-02	2.22E-01
provider_proivate	5.75E-01	1.43E-01	spend	-8.15E-04	2.32E-01
dist	1.37E-01	1.64E-01	female_calves	3.80E-01	2.40E-01
purchase_neighbor	-4.05E-01	1.66E-01	study_sitesouth_lift	2.42E-01	2.41E-01
prf_byr6	4.48E-01	1.73E-01	src_water_river	-2.22E-01	2.47E-01
tot_no_chils	4.47E-02	1.82E-01	freq_tctrl_num	-9.56E-02	2.52E-01
prf_byr2	-6.82E-01	1.88E-01	no_ctl_ownd_local	-3.77E-01	2.62E-01
haifers	-2.25E-01	1.92E-01	total_landsize	2.58E-02	2.65E-01
total_inc_crops	4.82E-08	1.96E-01	dist	-9.27E-02	2.81E-01
sys_rn_num4	3.85E-01	2.28E-01	provider_proivate	-2.35E-01	2.88E-01
rec_typs_record_sales	8.52E-01	2.44E-01	cows	3.30E-01	3.07E-01
mlk_prod_gehee	-1.75E+00	2.47E-01	dew_tms_num	-8.09E-02	3.16E-01
tt_market	5.30E-02	2.68E-01	ctl_id_notchning	-6.85E-01	3.18E-01
rec_usg_records_management	5.77E-01	2.72E-01	src_water_pond	4.80E-01	3.28E-01

	Tanzania			**Uganda**	
	Estimate	**Pr(>|t|)**		**Estimate**	**Pr(>|t|)**
brd_mthd_clvdnum	4.97E+00	4.38E-24	brd_mthd_clvdnum	5.11E+00	1.04E-23
provider_individual	-3.71E+00	3.63E-08	no_tms_ai_srvc	1.45E+00	2.16E-17
src_water_river	1.68E+00	1.39E-05	Factor1	7.47E-01	3.06E-04
provider_self	-3.38E+00	5.39E-05	provider_cooperation	3.38E+00	3.52E-04
ave_cst	5.75E-05	1.46E-04	rec_usg_asked_extension_officer	-2.28E+00	1.54E-03
dist	-8.99E-01	4.41E-04	rec_usg_self_evauation	2.26E+00	1.60E-03
no_tms_ai_srvc	4.78E-01	1.41E-03	food_crops	2.93E-02	2.73E-03
rec_usg_asked_extension_officer	-1.30E+00	4.10E-03	purchase_feed_suppplier	-4.60E+00	4.13E-03
rec_usg_records_identity	1.03E+00	5.09E-03	freq_tctrl_num	3.28E-01	4.14E-03
status_num	-7.09E-01	5.80E-03	rec_typs_record_growth	5.52E+00	1.67E-02
crop_res_mont	1.19E-01	7.99E-03	ave_cst	1.07E-05	2.38E-02
dew_tms_num	2.90E-01	9.63E-03	fnd_prf_mthd_num	2.29E+00	2.56E-02
freq_num	-2.95E-01	1.55E-02	immature_males	-6.86E-01	2.61E-02
rec_typs_record_feeding	1.02E+00	1.92E-02	kp_rcrds_num	1.92E+00	2.72E-02
study_sitembeya	-1.29E+00	2.00E-02	purchase_neighbor	-1.78E+00	2.91E-02
rec_usg_records_management	-7.55E-01	2.12E-02	hhh_yr_sch	-6.56E-02	4.39E-02
purchase_feed_local_trader	-1.12E+00	2.19E-02	no_ai_bull	-5.36E-01	4.63E-02
vac_avail_num	-6.25E-01	2.36E-02	src_residue_purchased	1.25E+00	6.94E-02
purchase_feed_suppplier	-2.34E+00	2.58E-02	mlk_prod_gehee	-4.62E+00	7.17E-02
src_water_tap	7.27E-01	4.46E-02	ctl_id_tags	-1.24E+00	7.95E-02
study_siteiringa	-1.25E+00	5.18E-02	provider_individual	-3.65E+00	8.58E-02
fodder_pur_mont	9.84E-02	5.62E-02	src_water_pond	8.29E-01	1.15E-01
study_sitetanga	9.10E-01	5.81E-02	src_water_borehole	-8.62E-01	1.28E-01
fnd_prf_mthd_num	-5.53E-01	6.17E-02	sys_dr_num	9.62E-01	1.39E-01
study_sitenjombe	-1.25E+00	6.22E-02	res_mlk	-2.53E-01	1.51E-01
kp_rcrds_num	9.61E-01	7.82E-02	crop_res_mont	8.22E-02	1.57E-01

rec_typs_record_milk	5.19E-01	7.89E-02	rec_typs_record_culving	-1.29E+00	1.81E-01
castrated_adult_males	-9.13E-01	9.18E-02	rec_usg_records_identity	9.71E-01	1.86E-01
ctl_id_markers	3.06E+00	9.38E-02	fodder_crops	1.97E-02	1.91E-01
freq_tctrl_num	9.28E-02	1.01E-01	src_water_river	-1.26E+00	2.06E-01
workinghrs	8.03E-02	1.11E-01	ctl_id_id_name	-1.06E+00	2.47E-01
src_scavenged_peoples_farm	3.65E-01	1.41E-01	wh_local_trader	1.70E+00	2.71E-01
wh_feed_supplier	-6.74E-01	1.93E-01	src_residue_own_farm	5.57E-01	2.84E-01
src_water_borehole	5.09E-01	1.96E-01	time_ft	-5.54E-01	2.90E-01
prf_byr	-8.90E-02	1.97E-01	purchase_feed_local_trader	-7.32E-01	3.32E-01
src_water_rain	-1.43E+00	2.01E-01	provider_self	-3.81E+00	3.42E-01
rec_usg_records_sales	7.57E-01	2.10E-01	rec_typs_record_milk	6.41E-01	3.46E-01
rec_usg_tracerbility	-4.12E-01	2.38E-01	sys_rn_num	-5.98E-01	3.60E-01
res_mlk	2.59E-02	2.52E-01	hhh_exp_num	9.96E-02	3.66E-01
hhh_exp_num	6.72E-02	2.59E-01	provider_proivate	8.27E-01	3.79E-01
fodder_one	-2.27E-01	3.15E-01	totalLlabour	-1.49E-01	3.81E-01
provider_proivate	6.30E-01	3.17E-01	no_ctl_ownd_local	1.38E-01	3.84E-01
avail_num	3.55E-01	3.40E-01	rec_typs_record_health	5.64E-01	4.53E-01
cst	-9.31E-02	3.42E-01	rec_typs_record_breeding	-5.41E-01	4.66E-01
total_landsize	8.75E-02	3.61E-01	rec_usg_tracerbility	-6.03E-01	4.77E-01
src_residue_purchased	6.47E-01	3.64E-01	status_num	-3.61E-01	5.13E-01
totalLlabour	-2.38E-01	3.78E-01	provider_goverment	-5.66E-01	5.40E-01
trans_cost	1.65E-04	4.01E-01	rec_typs_record_sales	-6.56E-01	5.48E-01
ctl_id_id_name	4.87E-01	4.11E-01	prf_byr	-5.84E-02	5.56E-01
bulls	3.99E-01	4.21E-01	cash_crops	2.21E-02	5.62E-01



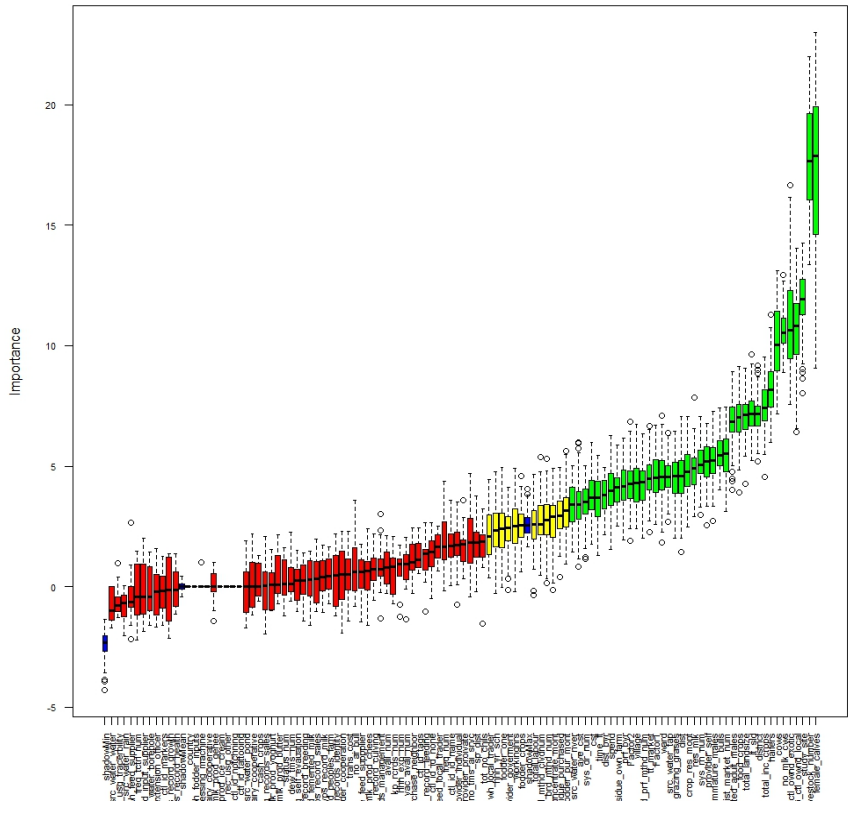
a) Ethiopia

b) Kenya

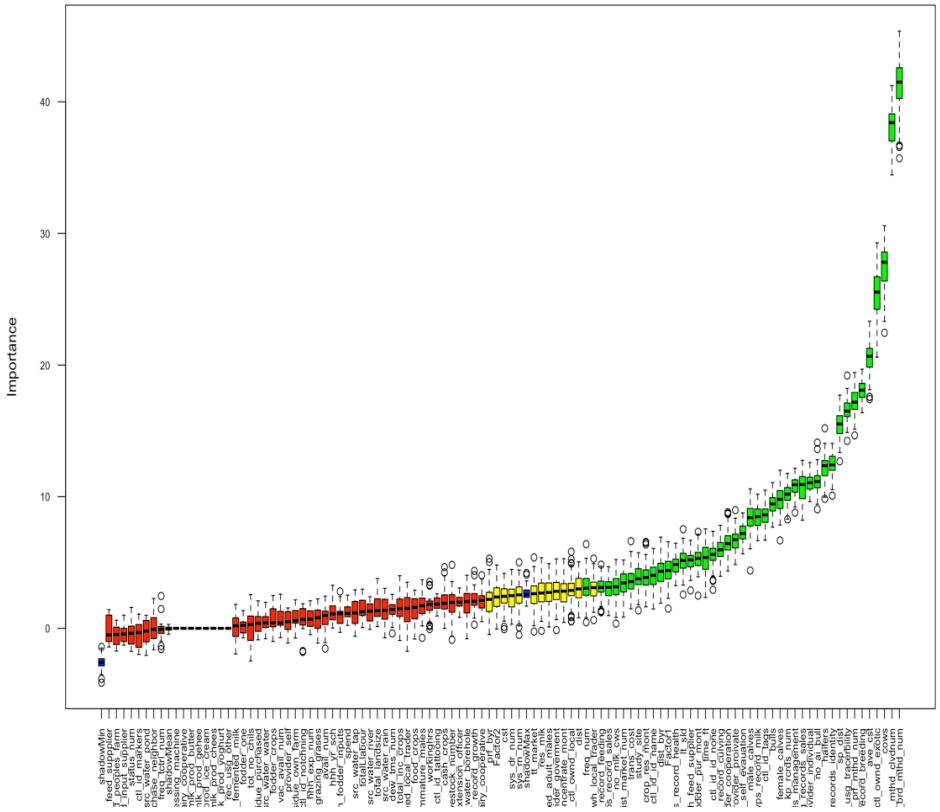
c) Tanzania

d) Uganda

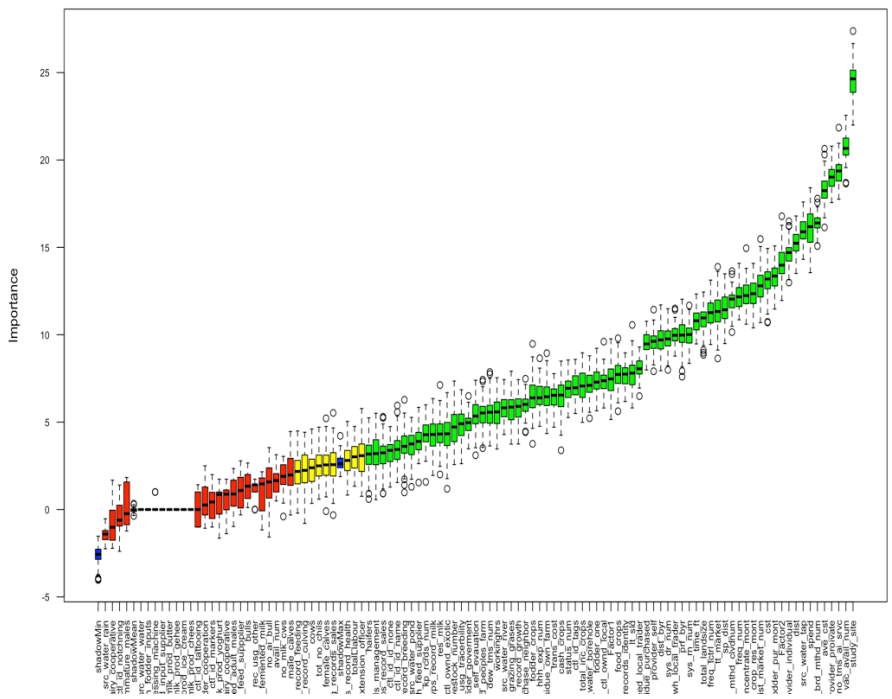
Figure 33: Variables selected by linear models to be used in developing prediction model to predict farmers decision in regard to breeding method in Ethiopia, Kenya, Tanzania and Uganda.



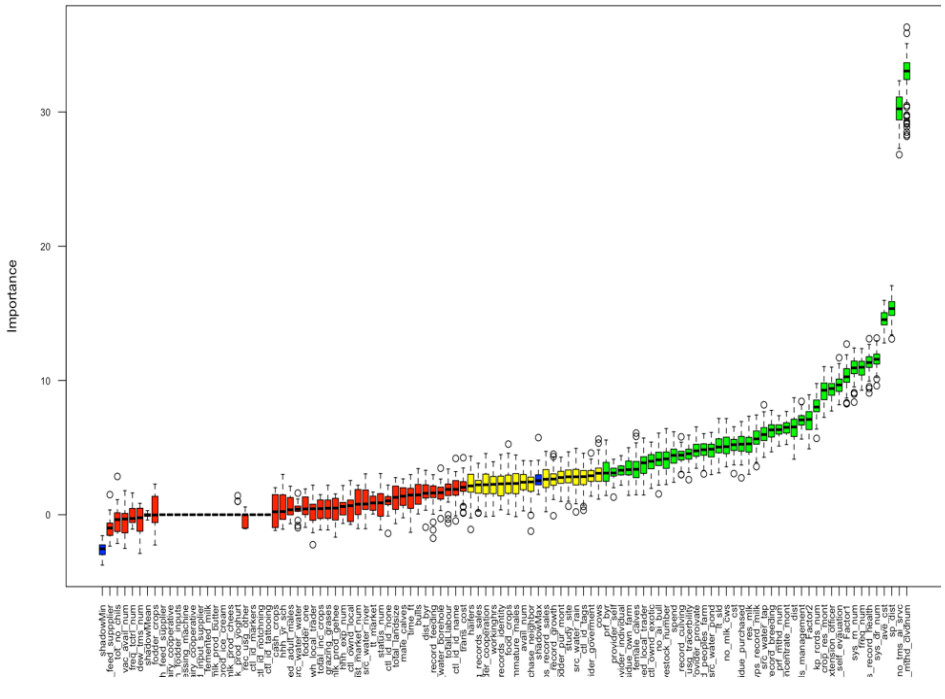
(a) Ethiopia



(b) Kenya

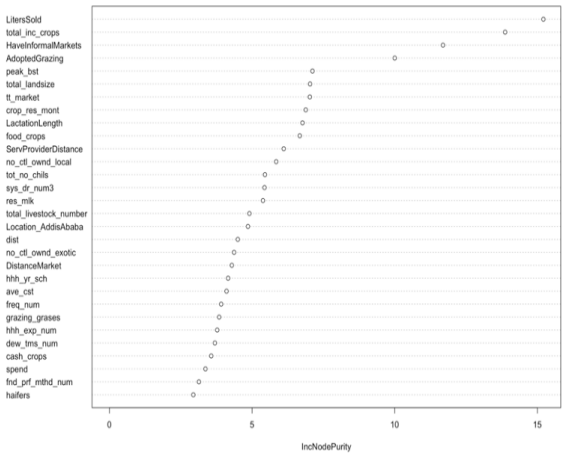


(c) Tanzania

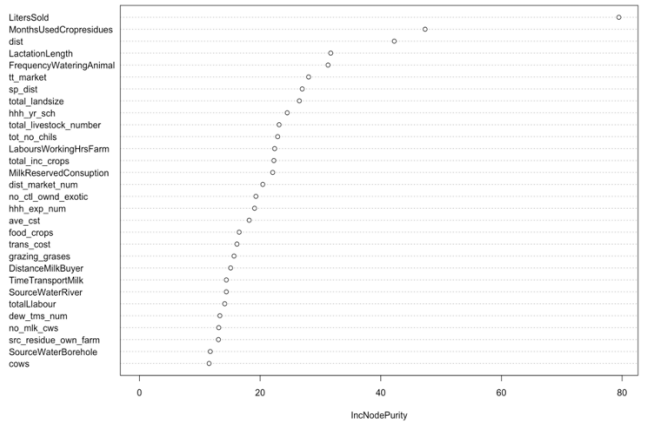


(d) Uganda

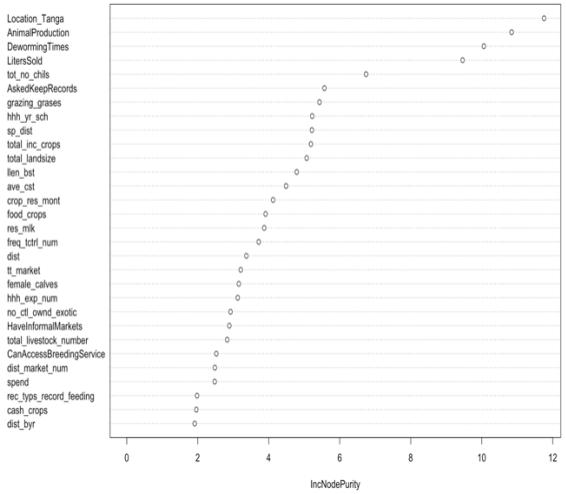
Figure 34: Variables selected by Boruta models to be used in developing prediction model to predict farmers decision in regard to breeding method in Ethiopia, Kenya, Tanzania and Uganda.



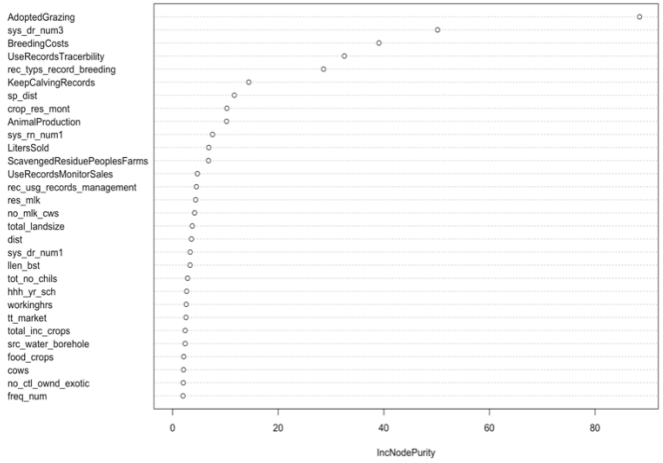
(a) Ethiopia



(b) Kenya

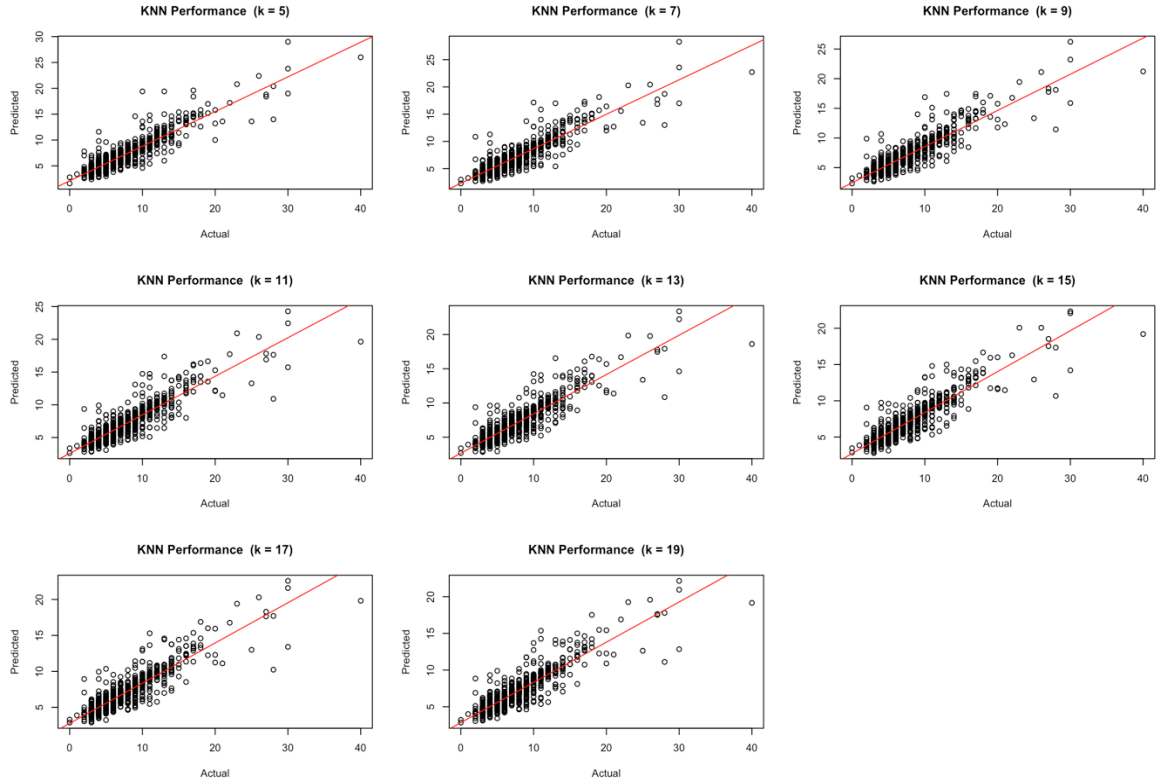


(c) Tanzania

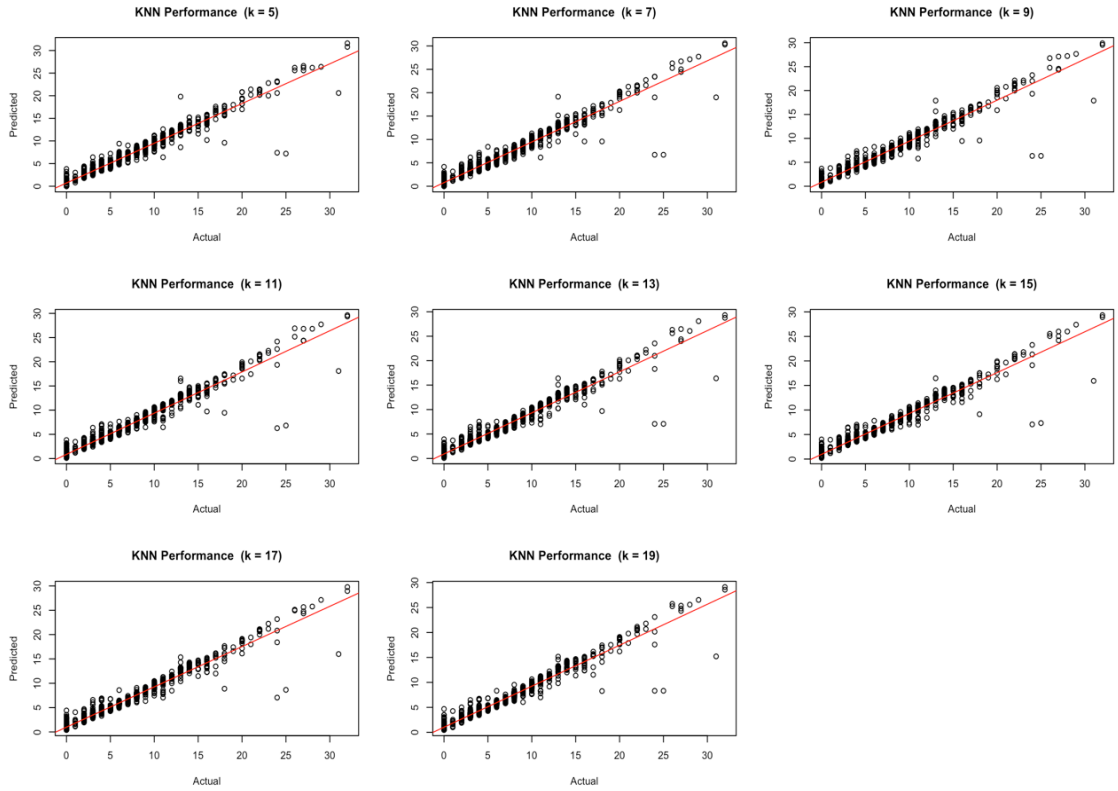


(d) Uganda

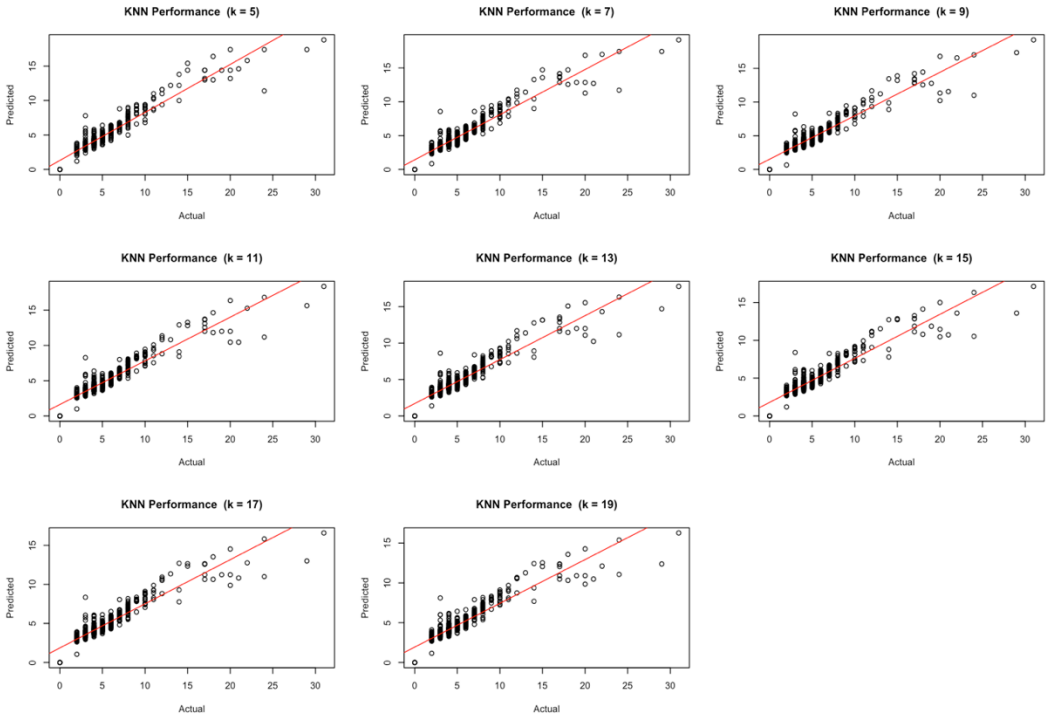
Figure 35: Variables selected by random forest models to be used in developing prediction model to predict farmers decision in regard to concentrate usage in Ethiopia, Kenya, Tanzania and Uganda



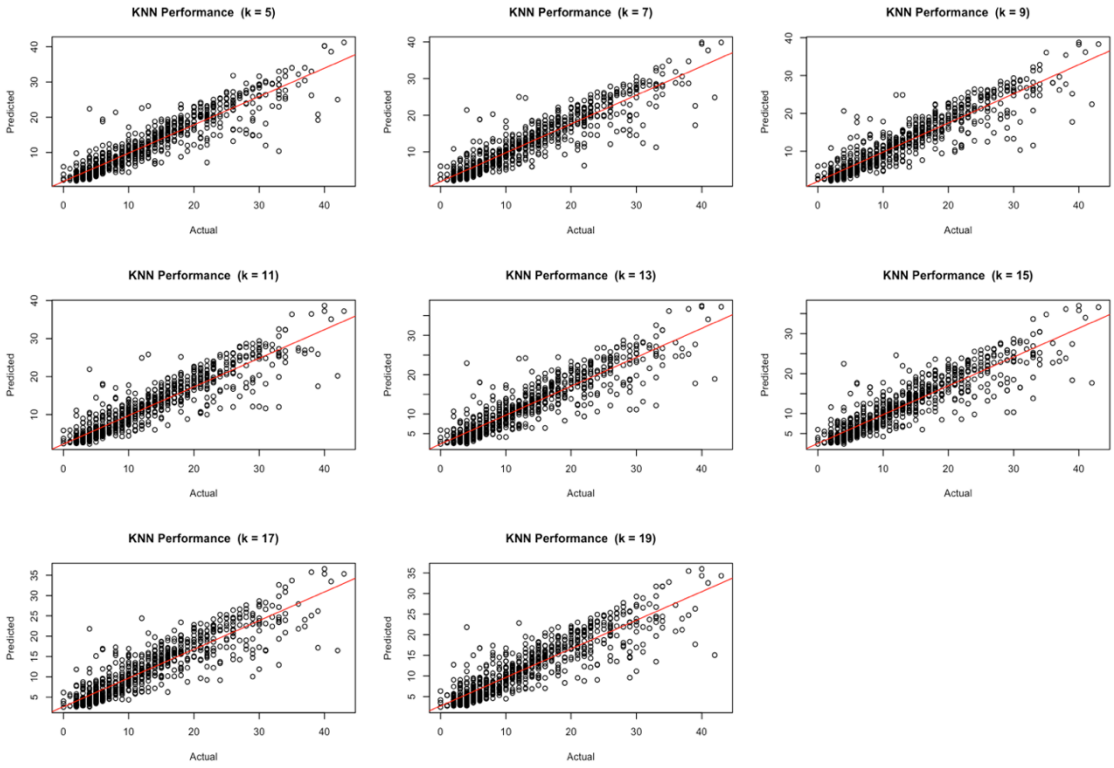
(a) Ethiopia



(b) Kenya

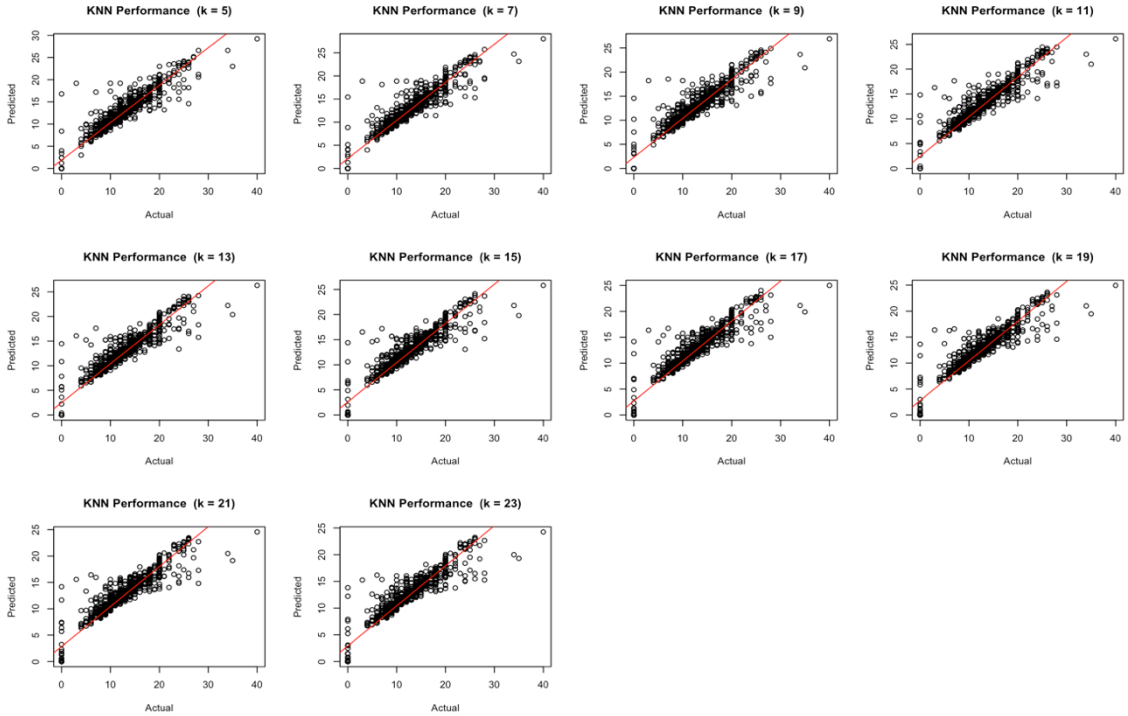


(c) Tanzania

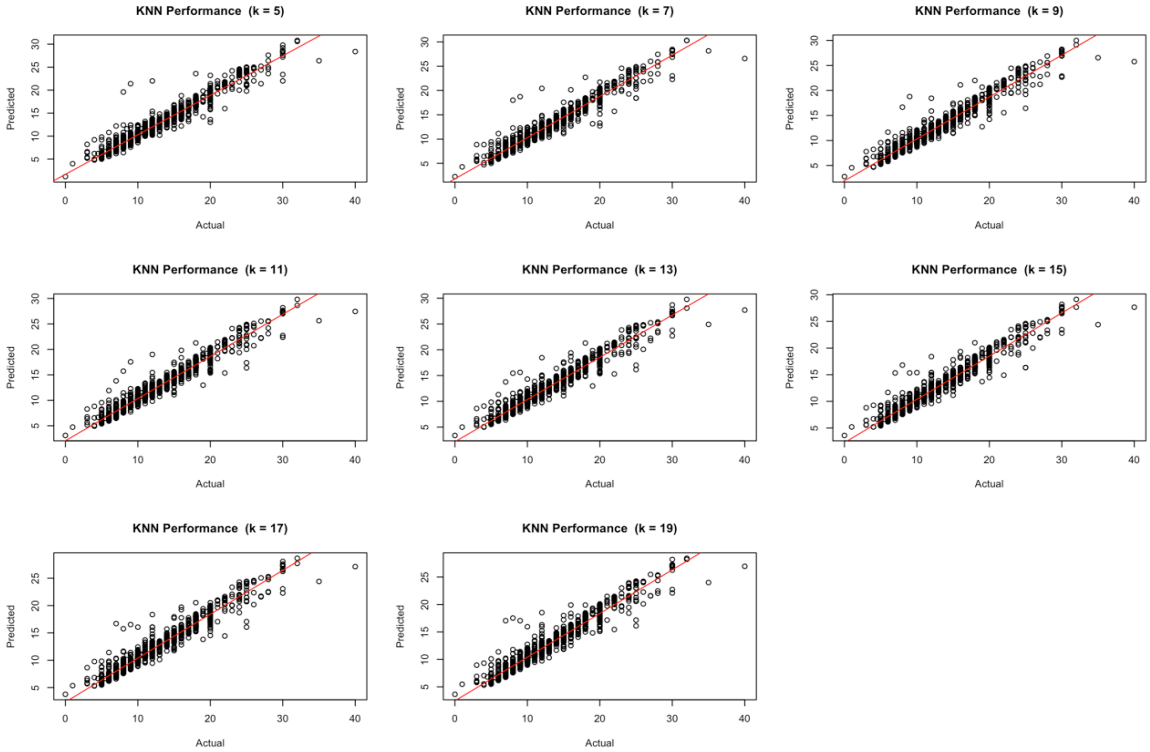


(d) Uganda

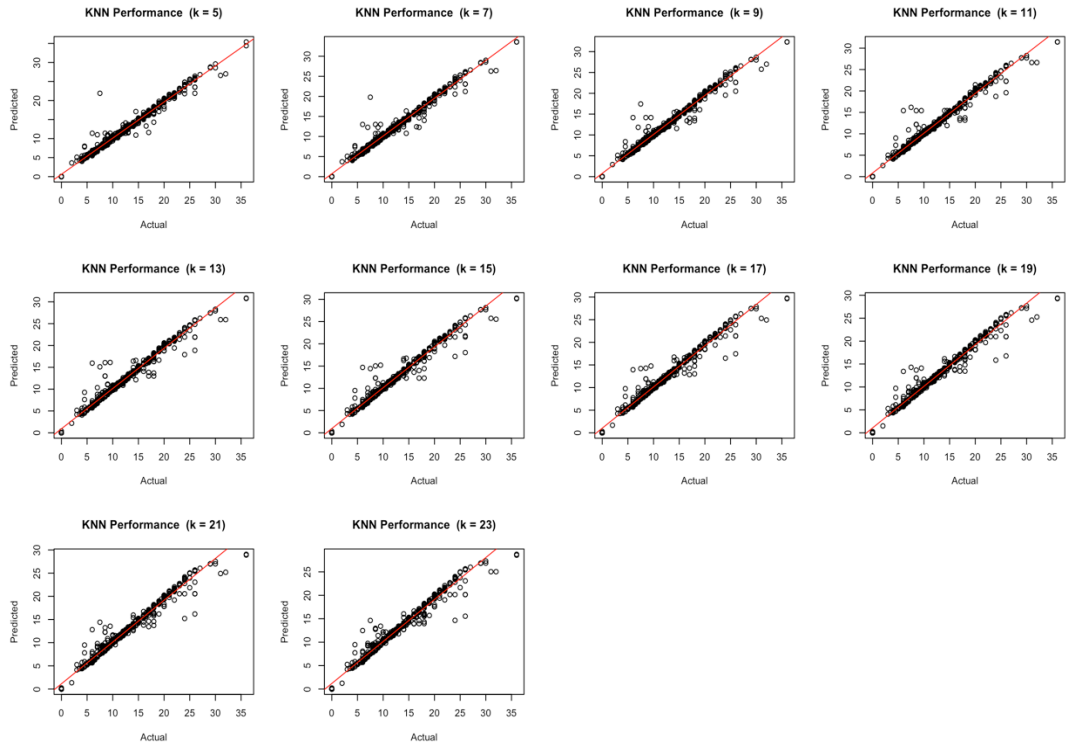
Figure 36: Performance for KNN model with different value of K for predicting the number of exotic animals to be kept on the farm in Ethiopia, Kenya, Tanzania and Uganda



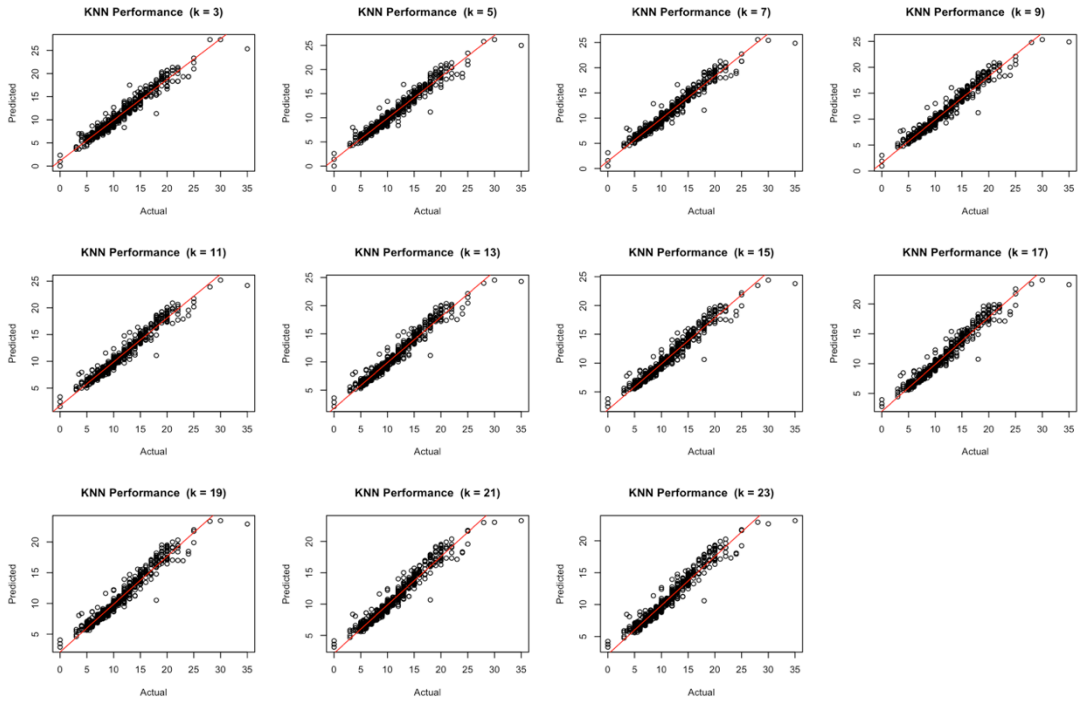
(a) Ethiopia



(b) Kenya



(c) Tanzania



(d) Uganda

Figure 37: Performance for KNN model with different value of K for predicting animal production in Ethiopia, Kenya, Tanzania and Uganda

Appendix 4: Questioner used to collect data

S1: General Identification
S1Q1: Please indicate the country
S1Q2: Please indicate the study site.
S1Q6: Please indicate the name of the district?
S1Q7: Please indicate the name of the ward?
S1Q8: Please indicate the name of the Village
S1Q9: What is the enumerator's name?
S1Q10: Capture the GPS Coordinates.
S1Q11: Name of the respondent
S1Q12: Gender of the respondent
S1Q13: Relationship of respondent to household head
S1Q14: Distance to the closest market center (km)
S1Q15: Time taken to closest market on foot (hrs)
S2: Business Owner/Household Head
S2Q1: What is the name of household head?
S2Q2: Please indicate the sex of the household head.
S2Q3: What is ${s2q1_hhh_name}'s experience in dairy farming?
S2Q4: Has ${s2q1_hhh_name} gone to school
S2Q5: How many years did ${s2q1_hhh_name} spend in school?
S2Q8: Can ${s2q1_hhh_name} read an official language?
S2Q9: Can ${s2q1_hhh_name} write an official language?
S2Q10: Who makes decisions about dairy activities?
S2Q11: How many members does this household have?
S2Q12: How many children aged between 10 years and 18 years are in this household?
S2Q13: How many children less than 10 years old are in this household?
S2Q14: How many children above 18 are in this household?
S2Q15: Total number of children in the household
S3: Agricultural Assets: value, ownership and access
S3: Land
S3Q1: Do you own/rent/squatter land
S3: Land ownership
S3: Type
S3Q2: How many plots of land do you (own/rent)
S3Q3: What is the main crop you cultivate in that plot?
S3Q4: What is the size of the plot (acres)
S3Q5: How much land from friends/relatives does the household use for free (acres)
S3Q6: How much land has been given out to friends/relative to use for free (acres)
S4Q1: What type of cattle do you own?

S4Q2: What are the breeds of the cattle do you own?
S4Q3: How many cattle (Based on their breeds) do you have?
S4Q4: What other livestock species does the household/business own?
S4: Other livestock species
S4Q5: What is the name of the other species?
S4Q6: How many animals per species do you own?
S4Q7: If you have cross bred or pure bred (grade) cows, how did you first acquire the cows
S4Q8: Have you stopped keeping grade cattle?
S4Q9: Indicate when you stopped keeping grade cattle
S4Q10: Reasons why you stopped keeping grade cattle
S4Q11: Have you purchased any cattle in the last one (1) year?
S4: Cattle types purchased
S4Q12: What cattle types were purchased?
S4Q13: What are the breeds of cattle you purchased?
S4Q14: How many cattle did you purchase?
S4Q15: What were the reasons for purchase?
S4Q16: What was the average price per animal?
S4Q17: From whom did you purchase the cattle?
S4: Cattle sold
S4Q18: Has the household sold any cattle in the last one year?
S4Q19: What cattle type were sold?
S4Q20: What are the breeds you sold?
S4Q21: What were the reasons for selling?
S4Q22: To whom did you sale the animals?
S4: Cattle dead
S4Q23: Has any cattle from your farm died in the last 1 year?
S4Q24: What cattle type died?
S4Q25: What are the breeds of animals that died?
S4Q27: How many animals died?
S4Q28: What were the causes of death?
S4Q29: Do you keep records for your cattle enterprise?
S4Q30: What types of records do you keep?
S4Q31: what do you use those records for?
S4Q32: what kind of cattle identification do you use on your farm?
S5: **Milk Production**
S5Q1: How many milking cows do you have?
S5Q2: What was the average daily milk production at calving for your best cow? (Litres/day)
S5Q3: What was the average daily milk production at calving for your worst cow? (Litres/day)
S5Q4: What was the peak production for your best cow? (Litres)
S5Q5: What was the peak production for your worst cow? (Litres)
S5Q6: What was average daily milk production at late lactation for your best cow? (Litres/day)
S5Q7: What was average daily milk production at late lactation for your worst cow? (Litres/day)
S5Q8: What is the average lactation length for your best cow? (Months)

S5Q9: What is the average lactation length for your worst cow? (months)
S5Q10: Did you sell milk yesterday?
S5Q11: How many liters of fresh milk did you sell?
S5: Milk sale
S5Q12: To whom did you sell the milk?
S5Q13: What was the amount of milk sold to buyers
S5Q14: Who is your most preferred buyer?
S5Q15: Please give reasons for your selection of preferred buyer?
S5Q16: Do you transport milk to the buyers?
S5Q17: What is the distance to the buyer? (km)
S5Q18: The average travel time on foot to the buyers (hrs)?
S5Q19: What is the cost of transport to buyers yesterday (Tshs)?
S5Q20: How long does it take to be paid after milk delivery?
S5Q21: Apart from payment what other services do you receive from the buyers?
S5Q22: How do you determine the price of milk?
S5Q23: Apart from fresh milk do you sell other milk products?
S5Q24: What other milk products do you sell
S5Q25: Usually how many liters of fresh milk do you reserve for household consumption per day?
Input use, costs and Technology adoption
S6: Feeding system
S6Q1: What is the main livestock feeding system used on your farm during the rainy season?
S6Q2: what is the main livestock management system used in your farm during dry season?
S6Q3: what are the reasons for the choice of those livestock management systems?
S7: Water for cattle
S7Q1: How frequently do you water your cattle?
S7Q2: What water source do you use for your cattle?
S7Q3: what is the distance to the watering point (Km)?
S7Q4: Is the mentioned water source available throughout the year?
S7Q5: who collects water most regularly?
S7Q6: Do you pay for the water
S7Q7: How much do you pay per liter?
S7Q8: How much do you spend per week on purchasing water?
S8: Grown fodder
S8Q1: Do you grow any fodder?
S8Q2: If yes, what type of fodder do you grow?
S8Q3: What is the area of land you have grown fodder (acres)?
S8Q4: How do you treat the fodder before feeding the cattle?
S8Q5: Who is responsible for the fodder grown?
S8Q6: If you are not growing fodder, what are the possible reasons for not growing fodder?

S9: Purchased fodder
S9Q1: Do you sometimes purchase fodder for your cattle?
S9Q2: In which month of the last 1 year did you purchase fodder?
S9Q3: What type of fodder was purchased?
S9Q4: What cattle types were fed on purchased fodder?
S9Q5: Indicate from whom you purchased the fodder?
S9Q6: Where do you store forage?
S9Q7: Please indicate the methods you use to conserve feed on farm.
S10: Crop residues
S10Q1: Do you use crop residues?
S10Q2: Which of the last 1 year did you use crop residues?
S10Q3: What type of crop residue did you use?
S10Q4: What cattle type were fed?
S10Q5: What was the source of residue?
S10Q6: If purchased where did you purchase?
S10Q7: Have you sold crop residues to other farmers?
S10Q9: What units do you use for crop residue?
S10Q10: Value of crop residue sold in the last 1 year (per unit)
S11: Concentrates
S11Q1: Do you use concentrates?
S11Q2: Which of the last six (6) months did you use concentrates?
S11Q3: what type of concentrate did you use?
S11Q4: What cattle type were fed with the concentrates?
S11Q5: where did you purchase.
S11Q6: Can you afford the feeds (supplements) that you need for your animals?
S11Q8: Do you feed your animals a total mixed ration?
S11Q9: What is the source of ration?
S11Q10: Number of ingredients in formulated on farm ration?
S11Q11: What are the ingredients.
S11Q12: Proportion of ingredient (Percentage)
S11Q13: On average, how much ration does each animal consume?
S12: Breeding Services and Expenses
S12Q1: Which cattle breeds are you familiar with?
S12Q2: Please rank the breeds below according to your preference.
S12Q3: 1st Rank is
S12Q4: 2nd Rank is
S12Q5: 3rd Rank is
S12Q6: Why do you prefer the three top ranked breeds?
S12Q7: Whenever you want to buy a cow or serve your cow, which traits do you look for?
S12Q8: Please rank the top 3 traits below according to your preference.
S12Q9: 1st Rank is

S12Q10: 2nd Rank is
S12Q11: 3rd Rank is
S12Q12: Why do you prefer the three top ranked traits?
S12Q13: Which breeds in your opinion provides the desired traits?
S12Q14: Which are your preferred breeding methods?
S12Q15: Please indicate reasons for preference of AI breeding methods.
S12Q16: Please indicate reasons for preference of bull breeding methods.
S12Q17: If you wanted to breed/serve your cow can you find and use AI services?
S12Q18: How many times have you used AI services in the last 1 year?
S12Q19: What are the reasons for not using AI services
S12Q20: What are the reasons for not using bull service
S12Q21: what is the average cost per service?
S12Q22: Who are the service providers that you can access for this type of service?
S12Q23: What is the distance (km) from your farm to the service providers of your preferred method?
S12Q24: What breeding method did you use for the cow that calved most recently?
S12Q25: What was the average number of AI services before conception for the cows?
S12Q26: What was the average number of bull services before conception for the most recently calved cow?
S12Q27: Do you import semen?
S12Q28: From which country do you buy your semen?
S12Q29: What breeds do you purchase?
S12Q30: What types of semen do you purchase?
S12Q31: How many doses of sexed semen do you buy?
S12Q32: How many doses of regular semen do you buy?
S12Q33: Do you use embryo transfer technology?
S12Q34: Where do you source your embryos?
S12Q35: What is the average cost of the service per pregnancy?
S12Q36: How do you time estrus (heat) for your cows?
S12Q37: How do you ensure timely service/insemination for your cows?
S13: Animal Health Services and expenses
S13Q1: How many times have you dewormed your animals in the last 1 year?
S13Q2: Which animal types did you deworm in the last 1 year
S13Q3: Who provided the service?
S13Q4: Is tick control (spraying/dipping) service available?
S13Q5: How many times did you spray/dip your cattle in the last 1 year
S13Q6: Type of cattle treated in last 1 year
S13Q7: Who provided the spraying/dipping service?
S13Q8: Is vaccination service available?
S13Q9: How many times did you vaccinate your animals in the last 1 year
S13Q10: What type of cattle were vaccinated in last 1 year
S13Q11: What did you vaccinate your animals against?
S13Q12: Who provided the vaccination service?

S13Q13: Have you treated cattle for disease in the last one year
S13Q14: How many times were you visited in the last 1 year?
S13Q15: Please indicate if you have undertaken any kind of training on dairy care and handling in the last 1 year.
S13Q16: Do you regularly undertake evaluation of genetic merit for your animals?
S13Q18: How often do you do genetic evaluation?
S13Q19: Who provides genetic evaluation service?
S13Q20: How much do you currently pay for genetic evaluation?
S13Q21: How much would you be willing to pay per HERD to access regular genetic evaluation of your herd?
S14: Labor use and expenses
S14: Monthly Labor
S14Q1: Do you have a Monthly paid laborer(s)?
S14Q2: How many laborers do you employ?
S14Q3: How many of your employed laborers are women?
S14Q4: How many of your employed laborers are men?
S14Q5: What are the average working hours per day?
S14Q6: What is the average monthly wage per worker?
S14Q7: What are the main activities the laborers are engaged in?
S14: Monthly labor
S14: Current activity
S14Q8: What are the hours of work per day dedicated to each activity on the farm?
S15: casual laborer
S15Q1: Have you employed any casual laborer(s) in the last 1 year?
S15Q2: What types of activities are casual laborers involved in?
S15Q3: How many females have you hired?
S15Q4: How many males have you hired?
S15Q5: How many hours are allocated per person?
S16: Household labor
S16Q1: Have you used household labor in the last 1 year?
S16Q2: What household labor have you used in the last one year?
S16Q3: How many household members do you use as household labor?
S16Q4: How many hours per day do each household labor work per day?
S16Q5: What is the frequency of their involvement?
S17: Participation in Farmer Group and Dairy Market Hub
S17: Farmer groups
S17Q1: Do any household member belong to a Farmer Group?
S17Q3: Who in the household is a member of a Farmer Group?
S17Q4: When did she/he join the group?
S17Q5: What is the name of the group that he/she belongs to?
S17Q6: What type of group?

S17Q7: What are the two main functions that this group performs for you?
S17Q8: Does this cooperative or group own a chilling plant?
S17Q9: Has the member bought shares in the chilling plant?
S17Q10: Does the member hold a position of responsibility in the group?
S17Q11: What position does the member hold in the group?
S17Q12: Gender of the member who holds a position in the group
S18: Credit: Access and Utilization
S18Q1: Has any member of your household received credit in the last1 year?
S18Q2: Which member of your household received credit in the last 1 year?
S18Q3: What was the Source of credit?
S18Q4: What were the reasons for credit?
S19: Household Income
S19: Crop income
S19Q1: Do you grow crops?
S19Q2: How many types of crops did you harvest in the last 1 year?
S19Q3: What unit do you use to measure your crops}?
S19Q4: What was the total output for each crop grown?
S19Q5: What was the quantity of crop sold?
S19Q6: What was the average price for each crop?
S19Q7: Did you rent land for these crops?
S19Q8: What was the cost of seeds (Tshs)?
S19Q9: What was the cost of fertilizer (Tshs)?
S19Q10: What was the cost of manure (Tshs)?
S19Q11: What was the cost of pesticides (Tshs)?
S19Q12: What was the cost of machinery (Tshs)?
S20: Income from cattle products (products other than milk) and services
S20Q1: Do you sell cattle products other than milk and other dairy products?
S20Q2: What type of dairy products other than milk do you sell?
S20Q3: What was units per each package?
S20Q4: What quantity of product did you sell in the last 1 year?
S20Q5: What was the average price per unit /package (Tshs)?
S20Q6: Do you sell cattle services?
S20Q7: What type of services do you sell?
S20Q8: How many services did you sell in the last 6 months?
S21Q1: Did you have any other income source(s) in the last 6 months?
S21Q2: If yes, what sources?